Universidad Autónoma de Madrid
Escuela politécnica superior

Trabajo Fin de Grado

ESTUDIO DE ANÁLISIS DE SUPERVIVENCIA

Doble Grado en Ingeniería Informática y Matemáticas

Cristina Pruenza García-Hinojosa
Mayo 2014

ESTUDIO DE ANÁLISIS DE SUPERVIVENCIA

AUTOR: Cristina Pruenza García-Hinojosa
TUTOR: Ana González Marcos

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Mayo 2014

Resumen

En la actualidad, miles de personas sufren o mueren de cáncer cada día y la comunidad médica todavía no puede hacer predicciones exactas con respecto al tiempo de supervivencia de un paciente afectado, sino que, se basan en sus observaciones y experiencias previas, llegando a dar pronósticos inexactos.

A partir de esta situación, surge el interés de realizar un acercamiento descriptivo y evolutivo para pacientes que sufren cáncer de mama, esperando que a partir de este estudio se puedan determinar los factores pronóstico de esta enfermedad, y así poder mejorar la calidad de la atención de dichos pacientes.

El objetivo de este trabajo consiste en desarrollar un análisis de supervivencia sobre pacientes diagnosticados de cáncer de mama, empleando ciertos métodos estadísticos. La supervivencia es evaluada mediante el estimador de Kaplan-Meier y la identificación de los factores pronóstico mediante el modelo de regresión de Cox.

A partir de los resultados obtenidos, se puede concluir que, la supervivencia a 5 años en cáncer de mama es del 70 %, a 10 años del 62 % frente a un 53 % para 15 años. Para la supervivencia a 5 años se obtienen como factores pronóstico el tamaño del tumor y la existencia de ganglios linfáticos afectados, los cuales también aparecen para 10 y 15 años. Además, para la supervivencia a 10 años se obtiene el gen *HER2* y, para 15 años, la edad de diagnóstico, los genes *HER2* y *STK15* y el receptor de estrógeno.

El análisis de estos factores permite determinar que, en general, la edad de diagnóstico, el tamaño del tumor, el número de ganglios linfáticos afectados y la presencia de un número elevado de copias del gen *HER2* afectan negativamente a la supervivencia. Mientras que, por el contrario, el receptor de estrógeno disminuye el riesgo de muerte.

Palabras Clave

Análisis de supervivencia, estimador de Kaplan-Meier, modelo de regresión de Cox, función de riesgo, censura, cáncer de mama.

Abstract

Thousands of people are currently suffering or dying of breast cancer and medical community is not yet able to calculate the exact survival time of an affected patient. The only knowledge available is based on previous observation and experience, often resulting in inexact prognosis.

It is of huge interest to design a descriptive and evolutionary approach for patients affected by breast cancer. In this paper we aim to determine the prognosis factors for cancer, to improve the quality of attention provided to affected patients.

The purpose of this study is to develop a survival analysis of patients diagnosed with breast cancer, using certain statistical methods. Survival rates have been calculated using the Kaplan-Meier estimator, and the Cox regression model has been applied to the most significant variables contributing to survival.

Based on the obtained results, this paper conclude that the 5-year survival rate for breast cancer was 70 %, the 10-year survival rate was 62 % and the 15-year was 53 %. The obtained prognostic and predictive factors for breast cancer in the 5-year survival were the tumour size and the existence of involved lymph nodes, which also appear for 10 and 15 years. Furthermore, an additional factor in the 10-year survival was *HER2* gen, and in the 15-year survival, the factors identified were the age of diagnosis, the *HER2* and *STK15* genes and the estrogen receptor.

The study of these factors helps to determine that, overall, the age of diagnosis, the tumour size, the number of involved lymph nodes and a high number of *HER2* genes, are prognostic factors which negatively affected the survival. However, the estrogen receptor reduces the risk of death.

Key words

Survival analysis, Kaplan-Meier estimator, Cox regression model, hazard function, censoring, breast cancer.

Agradecimientos

Me gustaría que estas líneas sirvieran para expresar mi agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización del presente trabajo. En especial a mi tutora, Ana, por toda su dedicación e interés mostrado hacia este estudio. Todo su apoyo y ayuda han sido esenciales para poder llegar hasta este punto.

Quería agradecer a Julia, así como a mis compañeros del IIC, que de un modo u otro me han transmitido su confianza y motivación.

A mi grupo de amigas, Lara, Julia y Rocío porque han sido muchos los momentos vividos a su lado y, sin duda alguna, su amistad es una de las mejores cosas que me llevo de la carrera. Su apoyo y sus consejos han sido imprescindibles para que hoy esté donde estoy.

Pero si hay algún motivo por el que haya sido capaz de finalizar estos estudios, es mi familia, en especial mis padres y mi hermano, por haber estado siempre apoyándome tanto en los momentos buenos y malos, y porque sin ellos nada de esto hubiese sido posible.

Tampoco quería olvidarme, de mi abuela, por todo su cariño e ilusión, y por haber sabido seguir adelante ella sola. Y Mamen, que ha estado siempre demostrándome su ilusión por verme terminar.

No puedo dejarme a la persona más especial, Andrés, su apoyo desde que le conocí en esta escuela, es el mejor regalo que me puedo llevar de estos cinco años.

A todos los demás, que de una forma u otra habéis sido parte de esta carrera, gracias.

Cristina Pruenza García-Hinojosa
Mayo 2014

Índice general

1. Introducción **1**

1.1. Motivación del proyecto 1

1.2. Objetivos y enfoque 2

1.3. Estructura del documento 2

2. Marco Estadístico Teórico **5**

2.1. Introducción 5

2.2. Conceptos básicos de análisis de supervivencia 5

2.2.1. Tiempo de supervivencia o tiempo de falla 5

2.2.2. Censura y truncamiento 5

2.2.3. Censura 6

2.2.4. Truncamiento 6

2.3. Modelo de supervivencia 7

2.3.1. Función de supervivencia 7

2.3.2. Función de riesgo (*Hazard Function*) 8

2.4. Estimación de la función de supervivencia 8

2.5. Comparación de funciones de supervivencia 9

2.6. Modelo de Regresión de Cox 10

2.6.1. Verosimilitud parcial 11

2.7. Contrastes de hipótesis 13

2.7.1. Test de razón de verosimilitud 13

2.7.2. Test de Wald 13

2.7.3. Test de puntajes (*score test*) 13

2.8. Residuos 14

2.8.1. Residuos de Cox-Snell 14

2.8.2. Residuos de martingala 15
2.8.3. Residuos de desvíos (*deviance*) 16
2.8.4. Residuos de puntajes (*score*) 17
2.8.5. Residuos Schoenfeld 17
2.8.6. Residuos *dfbeta* 18

3. Desarrollo **19**

3.1. Introducción 19
3.2. Lenguaje de programación 19
3.3. Biblioteca utilizada 20
3.4. Conjunto de datos 20
3.5. Método de selección de variables 20

4. Desarrollo Experimental **23**

4.1. Introducción 23
4.2. *Toy-Example* 23
4.2.1. Conjunto de datos 23
4.2.2. Estimación de la función de supervivencia 24
4.2.3. Comparación de funciones de supervivencia 25
4.2.4. Ajuste del modelo de regresión de Cox 26
4.2.5. Verificación del modelo de Cox 29
4.3. Estudio con datos reales 31
4.3.1. Conjunto de datos 31
4.3.2. Estudio 1: todas las covariables 33
4.3.3. Estudio 2: sin *Age_diagnosis* 50
4.3.4. Estudio 3: sin *NPI* 53

5. Conclusiones y trabajo futuro **61**

5.1. Conclusiones 61
5.2. Trabajo futuro 63

Glosario de acrónimos **65**

Bibliografía **66**

A. Detalle de la biblioteca utilizada **69**

Índice de figuras

2.1. Ejemplo de gráfica de residuos de Cox-Snell frente a los residuos r_{cs_i}. 15

2.2. Ejemplo de gráfica de residuos de *deviance*. 16

2.3. Ejemplo de gráfica de residuos de *deviance* frente a la puntuación de riesgo. . . . 17

4.1. Estimación de Kaplan-Meier de la función de supervivencia para datos de reincidencia. 25

4.2. Estimación de Kaplan-Meier de la función de supervivencia para presos con y sin ayuda económica. 26

4.3. Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de reincidencia. 29

4.4. Residuos *dfbeta* para covariables de datos de reincidencia. 30

4.5. Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 5 años con la covariable *NPI*. 36

4.6. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 5 años. 37

4.7. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 5 años. 37

4.8. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 5 años. 38

4.9. Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 10 años con las variables *Tumor_Size*, *Nodal_Status* y *Her2_Norm*. 41

4.10. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 10 años. 42

4.11. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 10 años. 43

4.12. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 10 años. 43

4.13. Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 15 años con las variables *Age_diagnosis*, *NPI*, *Her2_Norm*, *ER_Norm* y *stk15_Norm*. 46

4.14. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años. 47

4.15. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años. 48

4.16. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años. 49

4.17. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. 51

4.18. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. 52

4.19. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. 53

4.20. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*. 55

4.21. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*. 55

4.22. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*. 56

4.23. Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. 58

4.24. Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. 58

4.25. Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. 59

Índice de tablas

2.1. Tabla de contingencia para el contraste de igualdad de funciones de supervivencia en dos grupos en el instante t_i. 9

4.1. Resultado de contraste de igualdad de funciones de supervivencia para la covariable *fin*. 26

4.2. Parámetros del modelo de Cox para el estudio *toy-example*. 27

4.3. Resultado de contraste del modelo de Cox para datos de reincidencia. 29

4.4. Información sobre el conjunto de datos de cáncer de mama. 31

4.5. Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 5 años. 34

4.6. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 5 años. 35

4.7. Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 10 años. 39

4.8. Covariables significativas en la selección de variables para datos de cáncer de mama en el estudio de supervivencia a 10 años. 40

4.9. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 10 años. 40

4.10. Resultado de contraste del modelo de Cox para cáncer de mama en el estudio de supervivencia a 10 años. 41

4.11. Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 15 años. 44

4.12. Covariables significativas en la selección de variables para datos de cáncer de mama en el estudio de supervivencia a 15 años. 45

4.13. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años. 45

4.14. Resultado de contraste del modelo de Cox para cáncer de mama en el estudio de supervivencia a 15 años. 47

4.15. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. 50

4.16. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*. 54

4.17. Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. 57

1 Introducción

1.1. Motivación del proyecto

Hoy en día, en muchos campos de estudio es importante conocer el momento de ocurrencia de algún evento que se puede considerar de interés. Con esto, lo que se pretende es medir el tiempo que transcurre hasta que sucede un cierto evento. Por ejemplo, en ensayos clínicos, la ocurrencia del evento es el instante del fallecimiento. Esta medida no queda limitada solamente a términos de vida o muerte, sino a otras situaciones de diferente índole, dependiendo del contexto que se considere, por ejemplo, puede medir el tiempo de recurrencia, tiempo que dura la eficacia de una intervención o tiempo de un aprendizaje determinado. Por tanto, la supervivencia es una medida de tiempo a una respuesta, fallo, muerte, recaída o desarrollo de una determinada enfermedad o evento. Al periodo de tiempo que tarda en ocurrir el suceso de interés comúnmente se le llama *tiempo de supervivencia* o *tiempo de falla*.

El análisis de supervivencia es una forma de modelar el tiempo de falla utilizando información sobre eventos que han ocurrido con anterioridad en circunstancias similares. Para ello, se necesitarán ciertas nociones estadísticas que serán explicadas en el documento.

Actualmente, estas técnicas son una pieza fundamental y de interés en diversas aplicaciones de Ingeniería y Economía, así como en Ciencias Biológicas y de la Salud (investigaciones médicas).

Debido a la gran cantidad de información a manejar en el análisis de supervivencia, es necesario la utilización de un software. Pero no sólo esto, en algunos casos, la escasez de datos es la que ha motivado el desarrollo de los modelos de supervivencia utilizando información parcial de los tiempos de falla. En este caso, la utilización de un software permite obtener información precisa y confiable del modelo de supervivencia, razones que han motivado al uso del paquete estadístico **R**, ya que además facilita el manejo de datos.

Además, es cierto que muchos de los datos proporcionados por los estudios clínicos se expresan en términos de supervivencia, lo que motiva a la realización de un estudio de investigación sobre estos temas.

En la actualidad, miles de personas sufren o mueren de cáncer cada día y la comunidad médica todavía no puede hacer predicciones exactas con respecto al tiempo de supervivencia de un paciente afectado, sino que, se basan en sus observaciones y experiencias previas, llegando a dar pronósticos inconsistentes.

A partir de esta situación, surgió el interés de realizar un primer acercamiento descriptivo y evolutivo para pacientes que sufren cáncer de mama, esperando que a partir de este estudio se tenga mayor conocimiento de la población con la que se está trabajando y se puedan evaluar mejor las estrategias terapeúticas empleadas, para así poder mejorar la calidad de la atención de dichos pacientes.

1.2. Objetivos y enfoque

Una vez establecida la motivación de este trabajo, el objetivo fundamental será realizar un estudio descriptivo sobre pacientes diagnosticados con cáncer de mama en el Hospital Universitario de Yale, en EEUU [1].

Para abordar dicho objetivo, en primer lugar, se analizarán los distintos métodos existentes detallados en el marco estadístico teórico para el análisis de supervivencia. Además, se utilizará el paquete estadístico **R**, el cual proporcionará una serie de rutinas que facilitarán el estudio.

Con el fin de obtener los mejores resultados en el trabajo, los datos disponibles sobre cáncer de mama serán filtrados para obtener aquellos verdaderamente relevantes.

A partir de estos datos, se estimará la función de supervivencia y se diseñará un modelo estadístico basado en la regresión de Cox.

Finalmente, se evaluarán los resultados obtenidos del estudio con algunos resultados reales publicados.

1.3. Estructura del documento

La memoria de este trabajo consta de los siguientes apartados.

- **Capítulo 1**: Introducción y motivación del proyecto.
- **Capítulo 2**: Marco estadístico teórico. Estudio de los principales métodos de análisis de supervivencia.
- **Capítulo 3**: Descripción de herramientas, técnicas elegidas y algoritmos utilizados para el análisis estadístico.

- **Capítulo 4**: Experimentos realizados para la evaluación de los algoritmos implementados. Diferentes estudios sobre los datos de cáncer de mama y exposición de resultados.
- **Capítulo 5**: Conclusiones obtenidas tras el análisis de resultados y posibles líneas futuras de investigación.
- **Referencias, glosario y anexos.**

2

Marco Estadístico Teórico

2.1. Introducción

Para poder llevar a cabo un correcto análisis de supervivencia, ha sido necesario un previo estudio de ciertos conceptos estadísticos, los cuales se describen en esta sección del documento. Se introducen algunas definiciones básicas, que serán empleadas durante todo el trabajo y se describen técnicas para la estimación de la supervivencia y la verificación de sus supuestos.

2.2. Conceptos básicos de análisis de supervivencia

2.2.1. Tiempo de supervivencia o tiempo de falla

El estudio del análisis de supervivencia se centra en un grupo de individuos para los que se define un evento puntual. El tiempo de estudio (ya sean años, meses, semanas o días) de la ocurrencia de dicho evento comienza en un punto inicial de observación bien definido hasta un punto final establecido, momento en el que ocurre el evento o, por el contrario, finaliza el estudio. El evento puede ocurrir como mucho una vez en cualquier individuo.

Por tanto, el *tiempo de supervivencia* o *tiempo de falla* se define como el tiempo transcurrido desde la entrada de un individuo al estudio hasta el punto final.

2.2.2. Censura y truncamiento

En el análisis de supervivencia se pueden producir una serie de situaciones que complican la caracterización de los individuos que están dentro del estudio. La situación más favorable se da cuando es posible observar de manera exacta el tiempo T de aparición del suceso de interés. En esta situación se habla de datos no censurados.

Por otra parte, es habitual que algunos de los pacientes se pierdan a lo largo del estudio o tengan una entrada tardía, por lo que, no es posible realizar una observación completa de los tiempos de seguimiento. Esto es lo que se denomina *censura* y *truncamiento*.

2.2.3. Censura

Si el suceso de interés no ocurre durante el tiempo de observación del paciente, pero pasado este tiempo no se sabe cuándo ocurrirá el evento, es lo que se denomina *censura*. Por ejemplo, en un estudio de supervivencia después de un transplante, si el evento de interés es la muerte, puede pasar que un paciente deje de acudir a la consulta por un cambio de domicilio, perdiéndose su rastro a efectos de observar la muerte.

Existen varias categorías de censura, principalmente, *censura por la derecha*, *censura por la izquierda* y *censura por intervalos*. Una información más amplia de los diferentes mecanismos de censura puede ser consultada en las referencias [2] y [3].

- **Censura por la derecha**

  Este mecanismo de censura es el caso más común de datos incompletos y se caracteriza por el hecho de que, durante el tiempo de observación del individuo, no se produce el evento que se desea observar. La falta de datos puede darse por diversas razones, entre las cuales, puede ser que, hasta el momento de la finalización del estudio no ha ocurrido el evento (siempre y cuando el periodo de observación sea finito), el individuo abandone el estudio o se haya producido otro evento que imposibilite que el suceso a observar ocurra.

- **Censura por la izquierda**

  Este tipo de censura suele ser poco común en el análisis de supervivencia. Se produce cuando en la primera observación que se realiza sobre el individuo, el evento que se desea observar ya ha ocurrido.

- **Censura por intervalos**

  Éste es un tipo de censura que ocurre cuando sólo se sabe que al individuo le ocurre el evento de interés en un intervalo de tiempo, es decir, entre el instante t_i y un tiempo t_j.

En el presente trabajo, se trata con datos censurados por la derecha, y se asume que si el dato está censurado es porque no ocurre el evento de interés en el tiempo observado. Los datos de supervivencia se presentan en la forma (t_i, δ_i) donde t_i es el tiempo de observación y, $\delta_i = 0$ si la observación es censurada y $\delta_i = 1$ cuando el evento de interés ocurre.

2.2.4. Truncamiento

El *truncamiento* es otra característica que puede presentarse en algunos estudios de supervivencia, el cual se define como una condición que presentan ciertos sujetos en el estudio y el

investigador no puede considerar su existencia. Al igual que en la censura, existen dos tipos de truncamiento.

- **Truncamiento por la izquierda**

  Este tipo de truncamiento también es conocido como *entrada tardía al estudio*. Como su propio nombre indica, ocurre cuando los sujetos entran al estudio en un tiempo posterior tras la ocurrencia del evento de interés. Cabe destacar que, esto es opuesto a la censura por la izquierda, donde se tiene información parcial de individuos que presentan el evento de interés antes de su entrada al estudio, por lo que, estos individuos no serán incluidos.

- **Truncamiento por la derecha**

  En este caso, el truncamiento se da cuando sólo individuos que han presentado el evento son incluidos en el estudio y ningún sujeto que no haya presentado aún el evento será considerado. Un ejemplo de muestras que contienen este tipo de truncamiento son los estudios de mortalidad basados en registros de muerte.

2.3. Modelo de supervivencia

Para poder establecer un modelo de supervivencia es necesario conocer el término de *variable aleatoria*.

Una *variable aleatoria* puede concebirse como un valor numérico que está afectado por el azar. Dada una variable aleatoria, no es posible conocer con certeza el valor que ésta tomará al ser medida o determinada, aunque sí se conoce que existe una distribución de probabilidad asociada al conjunto de valores posibles.

A continuación, se detallan dos conceptos relativos a la supervivencia:

2.3.1. Función de supervivencia

La *función de supervivencia* se define como la probabilidad de que una persona sobreviva (no experimente el evento de interés) al menos hasta el tiempo t. Una definición más formal es la siguiente.

Definición 1 *Sea T una variable aleatoria positiva (no negativa) con una función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$. La función de supervivencia $S(t)$ es*

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u. \tag{2.1}$$

Cabe mencionar que, *S(t)* es una función no creciente, tal que

$$S(0) = 1 \text{ y } S(t) = 0 \text{ cuando } t \to \infty.$$

Esto es, la probabilidad de sobrevivir al tiempo cero es uno y la de sobrevivir un tiempo infinito es nula.

2.3.2. Función de riesgo (*Hazard Function*)

En el estudio de supervivencia, un concepto importante es la probabilidad de que a un individuo le ocurra el evento de interés en el siguiente instante de tiempo Δt, dado que ha sobrevivido hasta el tiempo t. Esto es la *función de riesgo*, que se define como

$$h(t) = lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.2}$$

Aplicando la ley de probabilidad condicional a la ecuación (2.2), se obtiene

$$h(t) = lim_{\Delta t \to 0} \frac{P((t < T \leq t + \Delta t) \cap (T > t))/P(T > t)}{\Delta t} = lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{P(T > t)\Delta t}. \tag{2.3}$$

Dado que

$$P(t < T \leq t + \Delta t) = \int_t^{t+\Delta t} f(u)\,\mathrm{d}u = F(t + \Delta t) - F(t). \tag{2.4}$$

Al sustituir (2.4) en (2.3) y aplicando la definición de derivada [1], se obtiene

$$h(t) = lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{P(T > t)} = \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}.$$

De manera que, la función de riesgo se define como

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.5}$$

Esta función también puede ser llamada *tasa de mortalidad*, ya que en la mayoría de los estudios se considera como evento de interés la muerte.
Además, existe otra función importante relacionada con la tasa de mortalidad que es la *función de riesgo acumulada*, la cual se define como sigue

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\log S(t). \tag{2.6}$$

Así, se puede obtener otra expresión para la función de supervivencia

$$S(t) = \exp[-H(t)]. \tag{2.7}$$

2.4. Estimación de la función de supervivencia

La presencia de datos censurados o truncados hace que la función de supervivencia no se pueda obtener directamente a través de argumentos probabilísticos, siendo necesario utilizar algunos estimadores. Existen diversas formas de estimar la función de supervivencia, pero el estimador que se utiliza en el presente trabajo es el *estimador de Kaplan-Meier*, ya que no es necesario trabajar con periodos de tiempo, sino que los mismos tiempos de observación van contribuyendo a la estimación de la función de supervivencia.

[1] La derivada de f en x es el límite del valor del cociente diferencial, es decir, $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$.

El *estimador de Kaplan-Meier* [4], que considera datos que pueden presentar censura, se define como

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{n(t_i) - d(t_i)}{n(t_i)} \tag{2.8}$$

donde $n(t_i)$ y $d(t_i)$ indican el número de individuos en riesgo y el número de eventos, respectivamente, ocurridos en el instante t_i.

La varianza del estimador de Kaplan-Meier es

$$V(\hat{S}_{KM}(t)) = \hat{S}^2_{KM}(t) \sum_{t_i \leq t} \frac{d(t_i)}{n(t_i)[n(t_i) - d(t_i)]}.$$

En el caso de muestras grandes, el estimador Kaplan-Meier, en un tiempo fijo t, se distribuye aproximadamente normal. Entonces, un intervalo de confianza al $(100(1-\alpha))\,\%$ de $\hat{S}_{KM}(t)$ está dado por

$$\hat{S}_{KM}(t) \pm z_{1-\alpha/2} SE(\hat{S}_{KM}(t))$$

donde $z_{1-\alpha/2}$ denota el percentil de una distribución normal estándar de dos colas al nivel de significancia $1-\alpha/2$, esto es, $P(Z < z_{1-\alpha/2}) = 1-\alpha/2$, con $Z \sim N(0,1)$. Y $SE(\hat{S}_{KM}(t))$ es el error estándar de la estimación de Kaplan-Meier (la raíz cuadrada de la varianza).
En el estudio, se utilizan intervalos de confianza del 95 %, de modo que, $\alpha = 0{,}05$, así

$$\hat{S}_{KM}(t) \pm z_{0{,}975} SE(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t) \pm 1{,}96\sqrt{V(\hat{S}_{KM}(t))}.$$

2.5. Comparación de funciones de supervivencia

Hay numerosas situaciones en las que es necesario comparar dos curvas de supervivencia. Una de las formas más sencillas de llevar a cabo esto, es mediante una visión gráfica.

Sin embargo, existen diversas maneras analíticas de realizar una comparación, como puede ser a través de contrastes de hipótesis basados en tablas de contingencia (Tabla 2.1). Donde los grupos 1 y 0 indican cada una de las funciones de supervivencia.

Tabla 2.1: Tabla de contingencia para el contraste de igualdad de funciones de supervivencia en dos grupos en el instante t_i, donde los grupos 1 y 0 indican cada una de las funciones.

Evento	**Grupo 1**	**Grupo 0**	**Total**
Ocurre	$d_1(t_i)$	$d_0(t_i)$	$d(t_i)$
No ocurre	$n_1(t_i) - d_1(t_i)$	$n_0(t_i) - d_0(t_i)$	$n(t_i) - d(t_i)$
En riesgo	$n_1(t_i)$	$n_0(t_i)$	$n(t_i)$

Para la comparación de estas funciones se construye una prueba de hipótesis, es decir, un procedimiento que permite evaluar hasta qué punto un conjunto de datos observados es consistente con una hipótesis particular, conocida como hipótesis nula H_0.

En este caso, se toma como H_0 la igualdad de las funciones. De manera que, para construir el estadístico de contraste basta con calcular:

1. El número esperado de ocurrencias de eventos para cada grupo. Por ejemplo, para el grupo 1 es
$$\hat{e}_1(t_i) = \frac{n_1(t_i)d(t_i)}{n(t_i)}.$$
2. La varianza estimada del número de ocurrencias de eventos para cada grupo, la cual está basada en la distribución hipergeométrica y para el grupo 1 se define como
$$\hat{V}(d_1(t_i)) = \frac{n_1(t_i)n_0(t_i)[n(t_i) - d(t_i)]}{n^2(t_i)[n(t_i) - 1]}.$$

Finalmente, el estadístico de contraste para el grupo 1 se define de la siguiente manera

$$Q = \frac{[\sum_{i=1}^{m} w_i(d_1(t_i) - \hat{e}_1(t_i))]^2}{\sum_{i=1}^{m} w_i^2 \hat{V}(d_1(t_i))}. \tag{2.9}$$

Siendo m el número de tiempos de ocurrencia de eventos en ambos grupos y w_i son los pesos, que varían dependiendo del test utilizado. En el presente trabajo, se utiliza el *test log-rank*, en el que se toma $w_i = 1$. Existen otros test diferentes, los cuales pueden consultarse en [5].

Cuando el número de ocurrencias de eventos es demasiado grande, Q se puede aproximar mediante una distribución Chi-cuadrado χ^2 de un grado de libertad, es decir, $p = P(\chi^2(1) \geq Q)$. Si el valor del p-value p es menor que α (en este estudio, $\alpha = 0{,}05$), se rechaza H_0, pues significa que ambas funciones de supervivencia son diferentes.

2.6. Modelo de Regresión de Cox

En muchos estudios médicos, se incluye información adicional sobre cada individuo, de la cual se cree que puede depender el tiempo de supervivencia e, incluso, puede ayudar a determinar el pronóstico de los pacientes. El modelo de regresión descrito en el presente capítulo, conocido como *modelo de regresión de Cox* o *modelo de riesgos proporcionales de Cox*, permite estimar la relación que hay entre un conjunto de variables explicativas fijas $X_1, X_2, \ldots, X_n$, también conocidas como *covariables*, y la respuesta o tiempo de supervivencia; o más bien, con la función de riesgo $h(t; X)$, que es la tasa instantánea del suceso de interés.

En el modelo de regresión de Cox (1972), la función de tasa de riesgo del tiempo de falla de un modelo con vector de covariables dadas por X está definida de la siguiente manera

$$h(t; X) = h_0(t) \exp(\beta^\top X) \tag{2.10}$$

donde $h_0(t)$ es la *función de riesgo base*; y $\beta^\top = (\beta_1, \beta_2, \ldots, \beta_n)$ es el vector de parámetros de la regresión.

El modelo de Cox se dice que es un modelo semi-paramétrico, debido a que incluye una parte paramétrica y otra parte no paramétrica.

1. La parte paramétrica, $\exp(\beta^\top X)$, llamada *función de riesgo relativo*, la cual está claramente especificada y describe los efectos relativos de los parámetros de regresión estimados sobre el riesgo.

2. La parte no paramétrica es $h_0(t)$, la función de riesgo base, que es una función arbitraria y no especificada.

Además, tal y como se ha mencionado anteriormente, el modelo de regresión de Cox también es conocido como modelo de riesgos proporcionales, ya que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante en el tiempo, es decir,

$$\frac{h(t;X_i)}{h(t;X_j)} = \frac{h_0(t)\exp(\beta^\top X_i)}{h_0(t)\exp(\beta^\top X_j)} = \frac{\exp(\beta^\top X_i)}{\exp(\beta^\top X_j)} = \exp(\beta^\top (X_i - X_j)) = cte. \tag{2.11}$$

El cociente expresado en la ecuación (2.11) es conocido como *razón de riesgos relativos.*

2.6.1. Verosimilitud parcial

El objetivo del modelo de Cox consiste en estimar los parámetros β en la expresión (2.10). Para ello, Cox introdujo un método de estimación sin ser necesario una especificación previa de la función de riesgo base $h_0(t)$. Dicho método es la verosimilitud que, en 1975 [6], se llamó *función de verosimilitud parcial.*

Equivalentemente, se puede considerar la función de riesgo acumulado base $H_0(t) = \int_0^t h_0(u)\,\mathrm{d}u$ o la función de supervivencia base $S_0(t) = \exp[-H_0(t)]$, en lugar de la función de riesgo base.

La función usual de verosimilitud para un conjunto de datos viene dada por la siguiente fórmula

$$L(\beta) = \prod_{i=1}^{n} [h_0(t_i)\exp(\beta^\top X_i)]^{\delta_i} \exp[-H_0(t_i)\exp(\beta^\top X_i)] \tag{2.12}$$

siendo δ_i el indicador de censura, que tal y como se ha descrito anteriormente en el documento, $\delta_i = 0$ si la observación es censurada o, por el contrario, $\delta_i = 1$. Utilizando la expresión (2.7), la función de verosimilitud se puede expresar como

$$L(\beta) = \prod_{i=1}^{n} [h_0(t_i)\exp(\beta^\top X_i)]^{\delta_i} S_0(t_i)^{\exp(\beta^\top X_i)}. \tag{2.13}$$

En el modelo de Cox no se supone una forma específica de $h_0(t)$, por lo que, no es posible emplear directamente el método estándar (2.13) para obtener un estimador del vector de parámetros β. Pese a esto, Cox propuso la siguiente función de verosimilitud para estimar β

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^\top X_i)}{\sum_{j\in R(t_i)} \exp(\beta^\top X_j)} \right]^{\delta_i} \tag{2.14}$$

donde $R(t_i) = \{j : t_j > t_i\}$ es el conjunto que contiene los sujetos en riesgo en el tiempo t_i, es decir, que están en observación y aún no han presentado el evento.

Puesto que, δ_i es el indicador de censura y será 0 cuando la observación esté censurada, la función de verosimilitud parcial quedará

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^\top X_i)}{\sum_{j\in R(t_i)} \exp(\beta^\top X_j)} \tag{2.15}$$

siendo i ahora los índices de los m tiempos de los eventos observados, $t_1, t_2, \ldots, t_m$.

La correspondiente función de *log-verosimilitud parcial* es

$$\log(L(\beta)) = \sum_{i=1}^{m} [\beta^\top X_i - \log(\sum_{j \in R(t_i)} \exp(\beta^\top X_j))]. \tag{2.16}$$

De modo que, la estimación de los parámetros β se obtiene maximizando la función de verosimilitud parcial (ecuación (2.15)) o, de forma equivalente, maximizando la función de log-verosimilitud parcial (ecuación (2.16)) para $\beta_1, \beta_2, \ldots, \beta_n$, sin necesidad de estimar la función de riesgo base $h_0(t)$.

Sin embargo, cuando los datos contienen tiempos observados empatados [2], la verosimilitud parcial (2.15) puede llevar un tiempo de computación considerable. Por esta razón, cuando se tienen datos con empates, se utilizan aproximaciones para la función de verosimilitud parcial.

Una aproximación de la verosimilitud parcial bastante utilizada fue sugerida por Breslow (1974) [7]. Esta aproximación considera que los d_i eventos al tiempo t_i son distintos y ocurren secuencialmente.

La aproximación viene dada por la siguiente fórmula

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^\top S_i)}{[\sum_{j \in R(t_i)} \exp(\beta^\top X_j)]^{d_i}} \tag{2.17}$$

y el logaritmo de la verosimilitud

$$\log(L(\beta)) = \sum_{i=1}^{m} [\beta^\top S_i - d_i \log(\sum_{j \in R(t_i)} \exp(\beta^\top X_j))] \tag{2.18}$$

donde $S_i = \sum_{j \in D(t_i)} X_j$, $D(t_i) = \{i_1, \ldots, i_d\}$ es el conjunto de etiquetas de los individuos que experimentan el evento a tiempo t_i y d_i es el número de eventos a tiempo t_i.

Una aproximación alternativa, sugerida por Efron (1977) [8], la cual se considera más exacta que la de Breslow, es

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^\top S_i)}{\prod_{k=1}^{d_i} [\sum_{j \in R(t_i)} \exp(\beta^\top X_j) - \frac{k-1}{d_k} \sum_{j \in D(t_i)} \exp(\beta^\top X_j)]} \tag{2.19}$$

y el logaritmo de la verosimilitud

$$\log(L(\beta)) = \sum_{i=1}^{m} [\beta^\top S_i - \sum_{k=1}^{d_i} \log(\sum_{j \in R(t_i)} \exp(\beta^\top X_j) - \frac{k-1}{d_k} \sum_{j \in D(t_i)} \exp(\beta^\top X_j))]. \tag{2.20}$$

[2] El caso de datos de supervivencia sin tiempos empatados es más realista que ocurra cuando la variable del tiempo de supervivencia tiene una distribución continua. En cambio, con distribuciones discretas del tiempo de supervivencia, éstas permiten la presencia de empates en los datos, dependiendo también de la escala de medición que se considere (semanas, meses, años).

2.7. Contrastes de hipótesis

Tras el ajuste del modelo de Cox, se ha de comprobar si las variables del modelo son significativas. Para ello, existen pruebas que se encargan de validar la hipótesis de que los parámetros del modelo de Cox son asintóticamente equivalentes [9]. En ellas, se considera el vector de parámetros estimados $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)^\top$ y la matriz de información evaluada en β, definida como

$$I(\beta) = -\frac{\partial^2 \log(L(\beta))}{\partial \beta_k^2}, \quad k = 1, \ldots, n.$$

A continuación, se describen tres pruebas principales, las cuales son asintóticamente equivalentes, aunque pueden diverger al considerar la muestra estadística.

2.7.1. Test de razón de verosimilitud

Esta prueba es la que presenta una mayor confiabilidad. Sea la hipótesis nula $H_0 : \beta = \beta_0$ para la cual la estadística de prueba es

$$X_{LR}^2 = 2[\log(L(\beta_0)) - \log(L(\hat{\beta}))]$$

donde los β_0 son los valores iniciales de los coeficientes y $\hat{\beta}$ es el ajuste obtenido por el modelo de Cox.

Esta prueba sigue la distribución de Chi-cuadrado con n grados de libertad si H_0 es cierta para muestras grandes.

2.7.2. Test de Wald

Este test es, quizás, el más natural debido a que se basa en la distribución asintóticamente normal. El estadístico de contraste se define mediante

$$X_W^2 = (\hat{\beta} - \beta_0)^\top I^{-1}(\hat{\beta})(\hat{\beta} - \beta_0)$$

donde $I(\beta)$ es la matriz de varianzas y covarianzas estimada.

Esta prueba tiene como hipótesis nula $H_0 : \beta = \beta_0$ y sigue la distribución de Chi-cuadrado con n grados de libertad si H_0 es cierta para muestras grandes.

2.7.3. Test de puntajes (*score test*)

Esta tercera prueba es la conocida como test de los puntajes.

Se define $U(\beta) = (U_1(\beta), U_2(\beta), \ldots, U_n(\beta))^\top$ como el vector de derivadas de la función de log-verosimilitud parcial, $\log(L(\beta))$. Para muestras grandes, cuando H_0 es cierta, $U(\beta)$ tiene distribución asintótica normal con vector cero por media y matriz de covarianzas dada por $I(\beta)$. Teniendo como hipótesis nula $H_0 : \beta = \beta_0$, la estadística de prueba está dada por

$$X_{SC}^2 = U(\beta_0)^\top I^{-1}(\beta_0) U(\beta_0).$$

La cual tiene una distribución Chi-cuadrado con n grados de libertad.

2.8. Residuos

Una de las ventajas del enfoque del análisis de supervivencia es la posibilidad de efectuar un análisis de residuos [10].

Un residuo es el valor calculado, para cada paciente, como la diferencia entre el valor de supervivencia observado y el valor estimado por la ecuación de regresión. Cuanto mayor es esa diferencia mayor será el valor del residuo, con su signo correspondiente.

El análisis de residuos en cualquier modelo estadístico sirve para verificar la adecuación del modelo ajustado por medio de inspección de gráficos. De manera que, los residuos en el modelo de Cox pueden ser utilizados para:

- Descubrir la forma funcional correcta de un predictor continuo.
- Identificar los sujetos que están pobremente predichos por el modelo.
- Identificar los puntos o individuos de influencia.
- Verificar el supuesto del modelo de regresión de Cox.

Dentro de este marco se estudian seis tipos de residuos: residuos de Cox-Snell, de martingala, de desvíos (*deviance*), de puntajes (*score*) y de Schoenfeld. De estos residuos pueden derivarse otros, como los *dfbetas*.

2.8.1. Residuos de Cox-Snell

Este tipo de residuos, desarrollados por *Cox & Snell* [11], sirven para evaluar el ajuste global del modelo planteado. Si el modelo de regresión de Cox definido por (2.10) es adecuado, entonces las estimaciones del tiempo de supervivencia del modelo planteado vienen dadas por un estimador de la función de supervivencia $\hat{S}_i(t)$, el cual debe ser muy similar al valor verdadero de $S_i(t)$.

Para evaluar esto, se calculan los residuos de Cox-Snell para los n individuos en estudio del siguiente modo

$$r_{cs_i} = \hat{H}_0(t_i)\exp(\hat{\beta}^\top X_i),\ i = 1, \ldots, n.$$

Donde $\hat{H}_0(t_i)$ es el estimador de la función de riesgo acumulado base de Breslow definido por

$$\hat{H}_0(s) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\exp(\hat{\beta}^\top X_i(s))}. \tag{2.21}$$

El cual se basa en un proceso de conteo $N_i \equiv \{N_i(t), t \geq 0\}$, que para el i-ésimo sujeto es el número de eventos observados hasta el tiempo t y donde $Y_i(t)$ es un proceso $0-1$ que indica si el i-ésimo sujeto está en riesgo en el tiempo t.

Un resultado importante demostrado por *Moeschberger* [9], *Cox & Snell* [11] y *Collet* [12], es que, si el modelo apropiado se ajusta bien a los datos, entonces los residuos r_{cs_i}, tendrán para cada i un valor de distribución exponencial con tasa de riesgo igual a la unidad.

Para probar si los residuos de Cox-Snell están o no aproximadamente distribuidos de forma exponencial, se construye su gráfico de residuos (Figura 2.1). La lógica de este método es sencilla. Si dichos residuos están distribuidos de forma exponencial, entonces una estimación de la tasa de riesgo basada en r_{cs_i} representada frente a los residuos r_{cs_i} debería tender a una línea recta que pasa por el origen con pendiente igual a la unidad. Es decir, el riesgo acumulado $H_r(r_{cs_i})$ frente a los residuos r_{cs_i}, debería ser aproximadamente una línea recta que pasa por el origen con pendiente igual a 1.

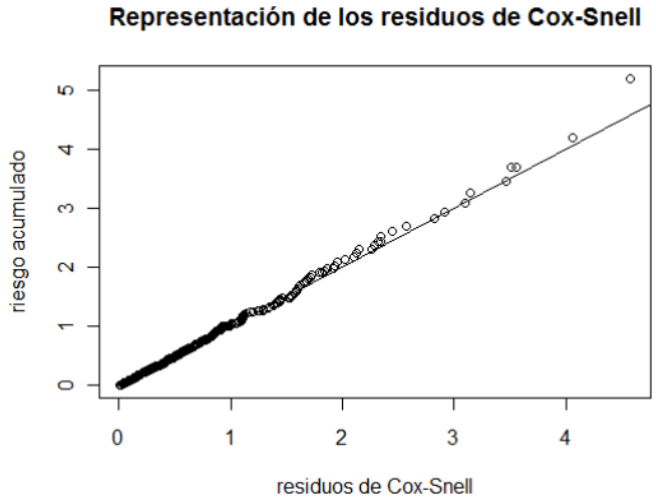


Figura 2.1: Ejemplo de gráfica de residuos de Cox-Snell, en la cual se representa el riesgo acumulado $H_r(r_{cs_i})$ frente a los residuos r_{cs_i}.

2.8.2. Residuos de martingala

Estos residuos están basados en la martingala de un proceso de conteo para el i-ésimo individuo, $M_i(t) = N_i(t) - E_i(t)$, esto es, la diferencia entre el proceso de conteo y la integral de la función de intensidad, tal y como sigue

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\beta^\top X_i(s)) h_0(s) ds,\ i = 1, \ldots, n.$$

Si se tiene β estimada por el estimador de máxima de verosimilitud parcial $\hat{\beta}$ y el estimador de riesgo acumulado base de Breslow definido en (2.21), el residuo de martingala se puede estimar de la siguiente forma

$$\hat{M}_i(t) = N_i(t) - \hat{E}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\hat{\beta}^\top X_i(s)) d\hat{H}_0(s).$$

Este residuo puede ser interpretado, para cada t, como la diferencia en el intervalo $[0, t]$ del número de eventos observados menos los esperados proporcionados por el modelo.

Por definición, el residuo martingala para un paciente censurado será negativo. Sin embargo, para pacientes en los que se produce el suceso de interés, el valor de los residuos puede ir desde $-\infty$ hasta 1. Si la muestra es grande, la suma de estos residuos es cero, no están correlacionados y el valor esperado es cero. Pero tienen el inconveniente de que no se distribuyen de forma simétrica en torno a cero, aunque el modelo sea correcto, lo que complica la interpretación de los gráficos. Por ello, se define otro tipo de residuos denominados *residuos de desvíos (deviance)*.

2.8.3. Residuos de desvíos (*deviance*)

Los residuos de *deviance* se obtienen mediante una normalización de los residuos martingala $\hat{M}_i$ y vienen dados por la expresión

$$d_i = signo(\hat{M}_i) * \sqrt{-\hat{M}_i - N_i \log\left(\frac{(N_i - \hat{M}_i)}{N_i}\right)}.$$

La transformación de los residuos martingala produce valores simétricos en torno a cero y, en este caso, el rango de valores de los residuos está comprendido entre $-\infty$ y $+\infty$.

Un residuo con un valor positivo grande corresponderá a pacientes que tienen un tiempo de supervivencia grande y los valores estimados por el modelo indican una supervivencia menor.

Este tipo de residuos se utiliza para la detección de valores atípicos (*outliers*). Para ello, se representan los residuos de *deviance* frente al índice de cada paciente, donde el índice corresponde al número de orden en el que el paciente ha sido registrado en el estudio. Se obtiene una gráfica como la de la Figura 2.2.

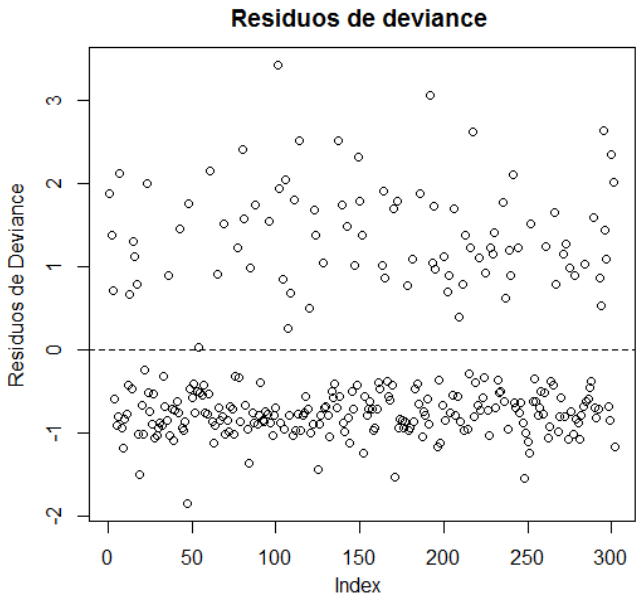


Figura 2.2: Ejemplo de gráfica de residuos de *deviance*, en la cual se representan los residuos frente al índice de cada paciente (número de orden en que ha sido registrado en el estudio).

Aunque en este ejemplo, Figura 2.2, no se aprecia ninguna anomalía evidente, en la zona superior aparecen algunas observaciones con residuos positivos grandes que convendría revisar. Para detectar dichas anomalías en el ajuste se representa, Figura 2.3, el valor de los residuos frente al resultado de calcular la puntuación de riesgo para el modelo estimado, es decir, el término $\beta^\top X$ donde cada β es el coeficiente estimado y cada X el valor de esa variable para cada paciente.

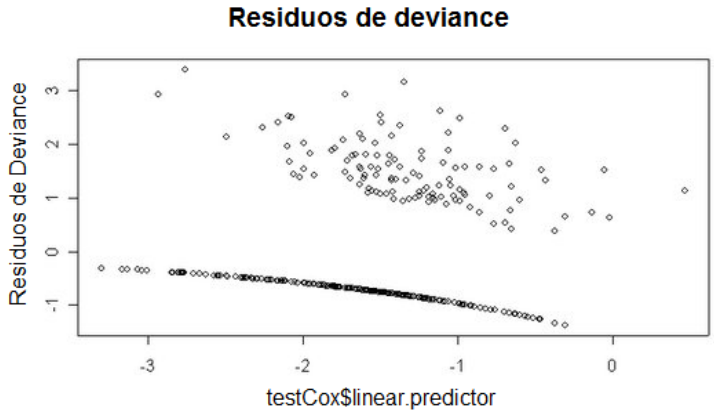


Figura 2.3: Ejemplo de gráfica de residuos de *deviance* frente a la puntuación de riesgo $\beta^\top X$ para cada paciente.

En la Figura 2.3 se muestra un ejemplo extremo de un modelo en el que se observa un patrón extraño en los residuos que revela que algo no está funcionando en el modelo estimado.

2.8.4. Residuos de puntajes (*score*)

Estos residuos se definen como

$$U_{ij} = U_{ij}(\hat{\beta}, \infty)$$

donde $U_{ij}(\beta, t)$, $j = 1, \ldots, n$ son las componentes del vector fila de longitud n obtenido a través del proceso de puntaje para el i-ésimo individuo, dado por

$$U_i(\beta) = \int_0^t [X_i(t) - \bar{X}(\beta, t)] dN_i(t)$$

donde $\bar{X}(\beta, t)$ es la media ponderada de las covariables sobre el riesgo establecido en el tiempo t.

Los residuos *score* se utilizan para verificar la influencia individual y para la estimación robusta de la varianza.

2.8.5. Residuos Schoenfeld

Los residuos de Schoenfeld se definen como la matriz

$$s_{ij}(\beta) = X_{ij}(t_i) - \bar{X}_j(\beta, t_i)$$

la cual contiene una fila por evento ocurrido y una columna por covariable, donde i y t_i son los individuos y el tiempo de ocurrencia del evento, respectivamente.

Los residuos de Schoenfeld son útiles para la verificación del supuesto de riesgos proporcionales en el modelo de Cox. Estos residuos, representados frente al tiempo observado de supervivencia deben repartirse aleatoriamente alrededor de cero, siempre que el modelo de Cox sea correcto.

2.8.6. Residuos *dfbeta*

Los residuos *dfbeta* sirven para determinar la influencia de cada observación en la estimación de los coeficientes de regresión.

Este residuo calcula el cambio aproximado en el j-ésimo coeficiente (es decir, la j-ésima covariable) si la observación i-ésima se elimina del conjunto de datos y se vuelve a estimar el modelo sin esta observación. Así para el paciente i el valor *dfbeta* correspondiente a la variable j es el siguiente.

$$dfbeta_i(\beta_j) = \beta_j - \beta_j(excluyendo\ i).$$

Los valores *dfbeta* pueden estandarizarse dividiendo por el error estándar del coeficiente correspondiente.

En su representación gráfica se suelen mostrar los valores de los residuos *dfbeta* estandarizados para cada covariable del modelo frente a los índices de paciente (número de orden). Si la supresión de una observación hace que el coeficiente incremente, el residuo *dfbeta* es negativo y viceversa.

3
Desarrollo

3.1. Introducción

En este capítulo se describen brevemente las herramientas y técnicas elegidas para la realización del análisis estadístico que se llevará a cabo posteriormente con un conjunto de datos reales.

Además de especificar el lenguaje de programación empleado, se realiza una descripción sobre el conjunto de los datos disponibles para el estudio y sobre los diferentes métodos para la selección de variables incluidas en el modelo.

3.2. Lenguaje de programación

En el desarrollo del análisis de supervivencia llevado a cabo en este trabajo, se hace uso del lenguaje **R**, por la gran cantidad de funciones y métodos estadísticos implementados en él y, por el hecho, de ser utilizado por la gran mayoría de la comunidad científica.

R es un lenguaje de programación de alto nivel diseñado, principalmente, para el estudio de grandes datos y estadísticos, siendo esto otro punto a favor para su elección, puesto que, abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a las más modernas técnicas de clasificación, pasando por los test clásicos y el análisis de series temporales.

El lenguaje **R** fue inicialmente escrito por Robert Gentleman y Ross Ihaka de la Universidad de Auckland en Nueva Zelanda. Actualmente, este lenguaje es el resultado de colaboraciones de ámbito mundial. El código de **R**[1] está disponible como software libre bajo las condiciones de la licencia GNU-GPL.

[1] `http://www.r-project.org`

3.3. Biblioteca utilizada

Existen diversas bibliotecas actualmente en **R** que permiten llevar a cabo un análisis de supervivencia, pero la más utilizada y la empleada en este trabajo, es *survival*. Dicha biblioteca es seleccionada por ser capaz de soportar datos que presentan censura.

Además, contiene numerosas rutinas utilizadas para el desarrollo de este estudio. Algunas de ellas son *Surv*, *survfit*, *coxph* o *cox.zph*. Sin embargo, para el cálculo de residuos se han implementado algunos algoritmos.

En el Anexo A se pueden consultar más detalles importantes sobre la biblioteca.

3.4. Conjunto de datos

Para la realización del estudio de supervivencia, se parte de datos clínicos previamente normalizados. De esta manera, se consigue que éstos sigan una distribución normal y no existan valores demasiados dispersos entre ellos.

El conjunto de datos se divide en dos:

- **Conjunto de entrenamiento (*training*)**

  Corresponde al conjunto de datos utilizado para la estimación de los parámetros del modelo. Dicho grupo está formado por un 70 % del conjunto inicial.

- **Conjunto de prueba (*test*)**

  En este conjunto no intervienen los datos de entrenamiento y, por tanto, se compone por el 30 % restante del conjunto de datos de partida.

  De modo que, una vez construido el modelo, se utiliza este conjunto para evaluarlo y verificar que éste se cumple para datos que no jugaron ningún papel en la selección del mismo.

3.5. Método de selección de variables

El principal objetivo a la hora de establecer un modelo es buscar aquel que contenga solamente los efectos principales, es decir, con las covariables que han resultado significativas de la regresión de Cox.

Para determinar cuáles de estas variables aportan más en el modelo y que, a su vez, no estén relacionadas, existen diferentes métodos de selección de variables, los cuales son:

- ***Forward* (hacia adelante)**

  En este método se inicia el proceso con un modelo vacío, sólo con el término independiente. Se ajusta un modelo con el método de máxima verosimilitud y se calcula el estadístico Chi-cuadrado con el p-value de incluir cada variable por separado.

A continuación, se selecciona el modelo con la variable más significativa, es decir, que tiene un p-value $p < 0{,}05$.

De nuevo, se ajusta el modelo con las variables seleccionadas y se calcula el p-value resultante de añadir cada una de las variables que no han sido seleccionadas por separado. Tras escoger el modelo con la variable más significativa, se repiten estos pasos hasta que no queden variables significativas por incluir.

- ***Backward* (hacia atrás)**

  Consiste en empezar con un modelo que contiene todas las variables candidatas e ir eliminando, una a una, cada covariable, a la vez que se calcula la pérdida de ajuste al eliminar.

  Se omite del modelo la variable menos significativa, esto es, aquella que mayor p-value tenga y se repiten estos pasos hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

- ***Stepwise***

  Este método es una combinación de los procesos *forward* y *backward*.

  Se puede comenzar, o bien, con el modelo vacío, o bien, con el modelo completo, pero en cada paso se exploran las variables incluidas por si deben salir y las no seleccionadas por si deben entrar en el modelo. Se repiten estos pasos hasta que todas las variables incluidas sean significativas y no entre ni salga ninguna más.

En este trabajo, se utiliza el método *backward*, por su facilidad a la hora de escoger la variable menos significativa y por permitir ir verificando que el proceso de eliminación es el adecuado.

Tal y como se ha mencionado anteriormente en el documento, se elige como nivel de significación de entrada para los estadísticos un p-value p menor que $\alpha = 0,05$.

4

Desarrollo Experimental

4.1. Introducción

En esta sección del documento, se procede a analizar los datos obtenidos tras llevar a cabo un estudio de supervivencia.

Para ello, en primer lugar, se realiza una serie de experimentos sobre un *toy-example*, de manera que sirva para comprobar que los algoritmos implementados y las funciones seleccionadas para llevar a cabo el estudio son las adecuadas y correctas, corrigiendo si fuese necesario algún posible error. A la vez que, se consigue cierto manejo sobre este tipo de datos.

En segundo lugar, se realiza el estudio de supervivencia sobre un conjunto de datos reales disponibles de cáncer de mama.

4.2. *Toy-Example*

4.2.1. Conjunto de datos

El análisis llevado a cabo se basa en un estudio sobre la reincidencia de presos durante el primer año después de ser liberados de las cárceles del estado de Maryland [2]. Se trata de un estudio piloto que sirve para ilustrar y verificar varios métodos de análisis de supervivencia.

El conjunto de datos consta de 432 presos, información recogida semanalmente, la cual se compone de las siguientes variables explicativas:

week Muestra la semana del primer arresto después de la liberación del preso. Si no ha sido un detenido reincidente, esta variable es censurada y toma el valor de 52.

arrest Es el indicador de censura, codificado como 1 si el preso fue arrestado durante el periodo en que ha estado en estudio y 0 en otro caso.

fin Indica si el preso recibe ayuda económica al salir de la cárcel. Toma el valor 1 si recibe ayuda y 0 en otro caso. En este estudio, se dio de manera aleatoria ayuda económica a la mitad de los presos.

age Es la edad, en años, que tiene el preso cuando es liberado.

race Indica el color de piel del preso. Toma el valor 1 si es negro o 0 en otro caso.

wexp Es la experiencia laboral, codificada como 1 si trabajaba a tiempo completo antes de ir a la cárcel o como 0 en otro caso.

mar Indica el estado civil. Vale 1 si está casado cuando el preso es liberado o 0 si no.

paro Variable que toma el valor 1 si el preso está en libertad condicional o 0 en otro caso.

prio Indica el número de encarcelaciones anteriores.

4.2.2. Estimación de la función de supervivencia

En primer lugar, para realizar el estudio sobre dicho conjunto de datos, se estima la función de supervivencia mediante el estimador de Kaplan-Meier, explicado en el apartado 2.4.

En la Figura 4.1, se muestra la función de supervivencia obtenida de la estimación para la reincidencia sobre el conjunto de datos de entrenamiento. Además, las líneas discontinuas representan los intervalos de confianza del 95 % alrededor de la curva estimada.

Se puede observar que, y como era de esperar, la probabilidad de supervivencia, en este caso, la no reincidencia, disminuye con el tiempo.

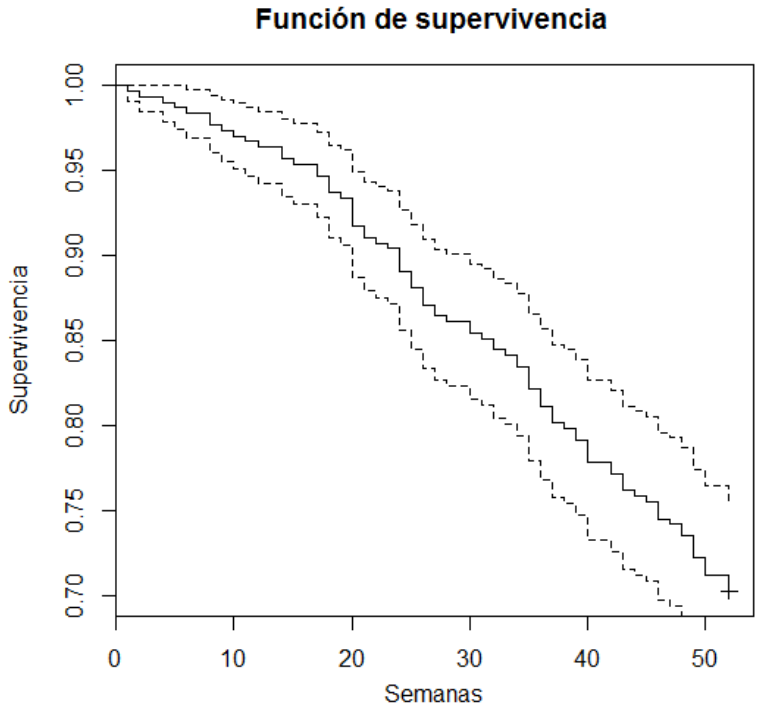


Figura 4.1: Estimación de Kaplan-Meier de la función de supervivencia para datos de reincidencia. Las líneas discontinuas representan los intervalos de confianza del 95 %.

4.2.3. Comparación de funciones de supervivencia

En este apartado, se comparan las funciones de supervivencia de presos que han recibido ayuda económica y los que no han recibido ningún tipo de ayuda.

Para ello, se construye un gráfico, Figura 4.2, donde se muestran las estimaciones de Kaplan-Meier para ambos casos.

Aparentemente, se observa que ambas funciones de supervivencia son distintas y que la probabilidad estimada de supervivencia, o lo que es lo mismo, de no reincidencia, es mayor en el grupo que recibe ayuda económica que en el grupo sin ayuda.

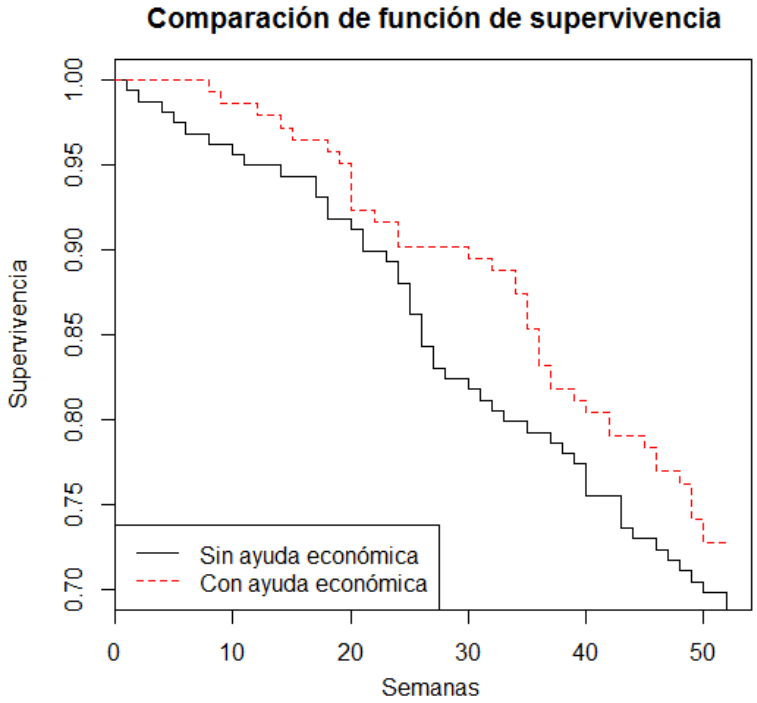


Figura 4.2: Estimación de Kaplan-Meier de la función de supervivencia para presos con y sin ayuda económica.

Además, haciendo uso de **R**, se comprueba que, efectivamente, existe diferencia entre ambas funciones. Obteniendo como resultado la información mostrada en la Tabla 4.1.

Tabla 4.1: Resultado de contraste de igualdad de funciones de supervivencia para la covariable *fin*. Aparece el número de sujetos en cada grupo, número observado y esperado de acontecimientos en cada grupo y el estadístico Chi-cuadrado para una prueba de igualdad.

	N	*Observados*	*Esperados*	$(O-E)^2/E$	$(O-E)^2/V$
fin=0	151	45	33,2	4,416	8,07
fin=1	151	24	35,8	3,87	8,07

Chi-cuadrado=8,1 con 1 grado de libertad, p=0,0045

En la Tabla 4.1, el estadístico de contraste *log-rank* (ecuación (2.9)) es $Q = 8{,}07$, lo cual se asocia con un p-value inferior a 0,05. Luego, la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5 %) se rechaza, lo que indica que hay diferencia entre las funciones de supervivencia del grupo de presos que reciben ayuda económica y el que no.

4.2.4. Ajuste del modelo de regresión de Cox

Puesto que se trata de un estudio piloto, se ajusta una sola vez el modelo de Cox en el que se incluyen todas las covariables, sin llevar a cabo el método de selección de las mismas. De manera que, los resultados del ajuste se incluyen en la Tabla 4.2.

Tabla 4.2: Parámetros del modelo de Cox para el estudio *toy-example*. Aparecen marcados los coeficientes significativos al 5 %.

n=302, número de eventos=90

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
fin	−0,172849	0,841264	0,213837	−0,808	0,418905
age	−0,080602	0,922560	0,026454	−3,047	0,002312 *
race	0,255544	1,291164	0,360845	0,708	0,478832
wexp	−0,153755	0,857482	0,230974	−0,666	0,505614
mar	−0,415430	0,660056	0,409924	−1,013	0,310854
paro	0,003469	1,003475	0,222124	0,016	0,987539
prio	0,117484	1,124664	0,031105	3,777	0,000159 *

Interpretación de los resultados

La Tabla 4.2 presenta información acerca de las pruebas locales para verificar que cada coeficiente es significativamente distinto de cero. Las columnas proporcionan información para cada covariable como sigue:

***coef*:** Valor del coeficiente de regresión estimado.

***exp(coef)*:** Función exponencial evaluada en el coeficiente de regresión estimado. Indica el cambio en la función de riesgo por cada unidad que se incremente la covariable asociada.

***se(coef)*:** Error estándar del coeficiente de regresión estimado.

***z*:** Corresponde al valor del estadístico, obtenido dividiendo el valor del coeficiente de regresión estimado entre el error estándar estimado.

***p*:** p-value proveniente de una distribución normal con media cero y varianza uno.

Además, se obtiene información correspondiente a probar la hipótesis nula de que el vector de variables del modelo son cero, es decir, $H_0 : \beta = \bar{0}$.

- Test de razón de verosimilitud

  La estadística de prueba, denotada anteriormente por X^2_{LR}, es 35,05 y con una distribución χ^2 con 7 grados de libertad, tiene un p-value $p = 1{,}093 \cdot 10^{-05}$, lo que significa que se rechaza la hipótesis nula.

- Test de Wald

  El test de Wald, denotado anteriormente por X^2_W, es 33,13 y con una distribución χ^2 con 7 grados de libertad, tiene un p-value $p = 2{,}504 \cdot 10^{-05}$. De igual modo, se rechaza la hipótesis nula.

- Test de puntajes

  La prueba de puntajes, denotada anteriormente por X^2_{SC}, es 35,28 y con una distribución χ^2 con 7 grados de libertad, se obtiene un p-value $p = 9{,}922 \cdot 10^{-06}$. Ocurre lo mismo que en los casos anteriores.

Para las tres pruebas anteriores se aprecia un p-value significativamente pequeño (inferior a 0,05), lo cual es evidencia de que, bajo las pruebas realizadas, los coeficientes del modelo son significativamente distintos de cero, y por tanto, se podrá considerar que el modelo tiene sentido para las variables explicativas consideradas.

Mediante los valores obtenidos en la Tabla 4.2, se puede verificar la significación de cada uno de los coeficientes correspondientes a las covariables. De manera que, los coeficientes de *age* y *prio* son significativos al 5 %, mientras que el resto de coeficientes no contribuyen significativamente en el modelo planteado y, por tanto, son variables candidatas a ser eliminadas.

Si se llevase a cabo el correspondiente método de selección de variables, en primer lugar, sería suprimida del modelo la covariable *paro*, por poseer el coeficiente menos significativo de entre las candidatas.

Además, los signos de los coeficientes son interpretables, de modo que, un valor positivo para un correspondiente β, significa que el riesgo aumenta con la presencia de la covariable, y un valor negativo, por el contrario, disminuye el riesgo. Así, la presencia de la covariable *fin* ($\beta = -0{,}172849$) disminuye el riesgo de ser arrestado, mientras que *prio* ($\beta = 0{,}117484$) aumenta.

Sin embargo, es más sencillo de interpretar la exponencial de los coeficientes β, *exp(coef)*, información obtenida directamente de la tabla anterior, en la cual se considera cada X_i como 1. Así, la función de riesgo base, $h_0(t)$, en la ecuación (2.10) aparece multiplicada por cada $\exp(\beta_i)$, permitiendo deducir el efecto que tiene cada coeficiente β sobre el riesgo.

A partir de esto, se puede decir que la aportación de ayuda económica hace que la reincidencia tenga un riesgo de 0,8413 veces el riesgo de reincidencia de los presos que no reciben ayuda económica, es decir, la ayuda económica disminuye el riesgo. En cuanto al número de encarcelaciones anteriores, covariable *prio*, aumenta el riesgo de una nueva detención en un factor de 1,1247.

En la Figura 4.3 se representa la función de supervivencia obtenida mediante el estimador de Kaplan-Meier y la obtenida por el modelo de Cox, donde se observa que la función de supervivencia estimada mediante Kaplan-Meier es sistemáticamente inferior al ajuste del modelo de Cox, pero se encuentra dentro del rango de los intervalos de confianza estimados.

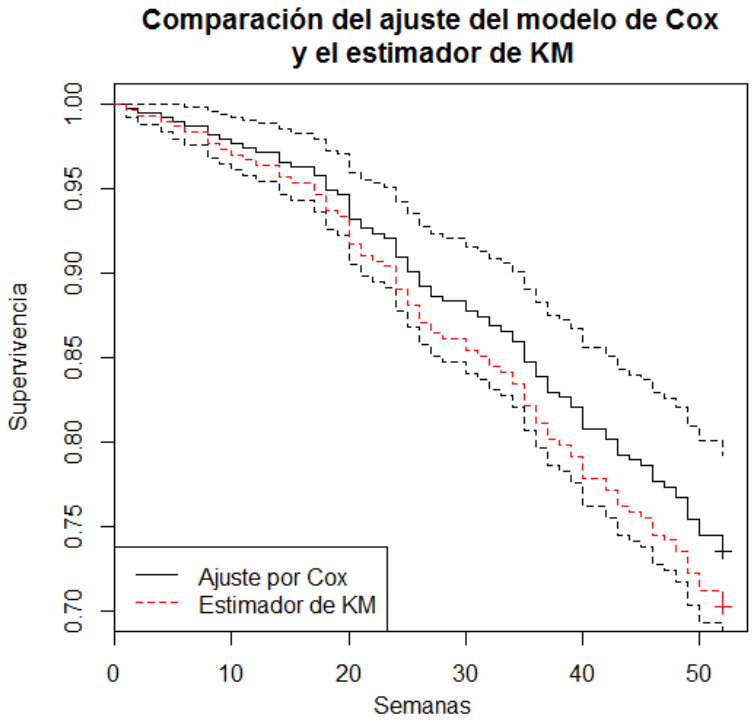


Figura 4.3: Comparación del ajuste del modelo de Cox, junto con los intervalos de confianza del 95 %, y el estimador de Kaplan-Meier para datos de reincidencia.

4.2.5. Verificación del modelo de Cox

1. **Contraste de hipótesis**

   Para verificar el modelo de regresión de Cox estimado se realiza un contraste de hipótesis, obteniendo la Tabla 4.3. En el que se considera como hipótesis nula el cumplimiento del supuesto de Cox, la cual se rechaza si el valor del p-value es inferior a 0,05.

Tabla 4.3: Resultado de contraste del modelo de Cox para datos de reincidencia, considerando como hipótesis nula el cumplimiento del supuesto de Cox. Para cada covariable se incluye el p-value resultante del contraste.

	p
fin	0,2332
age	0,0738
race	0,7992
wexp	0,0157
mar	0,9999
paro	0,3660
prio	0,8402
GLOBAL	0,1429

Puesto que, en este caso, no se ha llevado a cabo el método de selección de variables, la

covariable *wexp*, experiencia y dedicación laboral, viola el modelo de regresión de Cox, ya que tiene un p-value $p = 0{,}0157$.

2. **Residuos**

   Solamente se incluyen en este apartado los residuos *dfbeta*, puesto que, no se dispone de un modelo de Cox previamente ajustado y los resultados obtenidos en el resto de residuos son negativos para la aceptación del modelo.

   Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

   Con este tipo de residuo, se determina la influencia de cada observación en el modelo. De manera que, para cada covariable, se ha representado la observación (en orden de tiempo de falla registrado) por el cambio de escala aproximada del coeficiente, es decir, dividiendo por el error estándar, después de la eliminación de la observación del modelo.

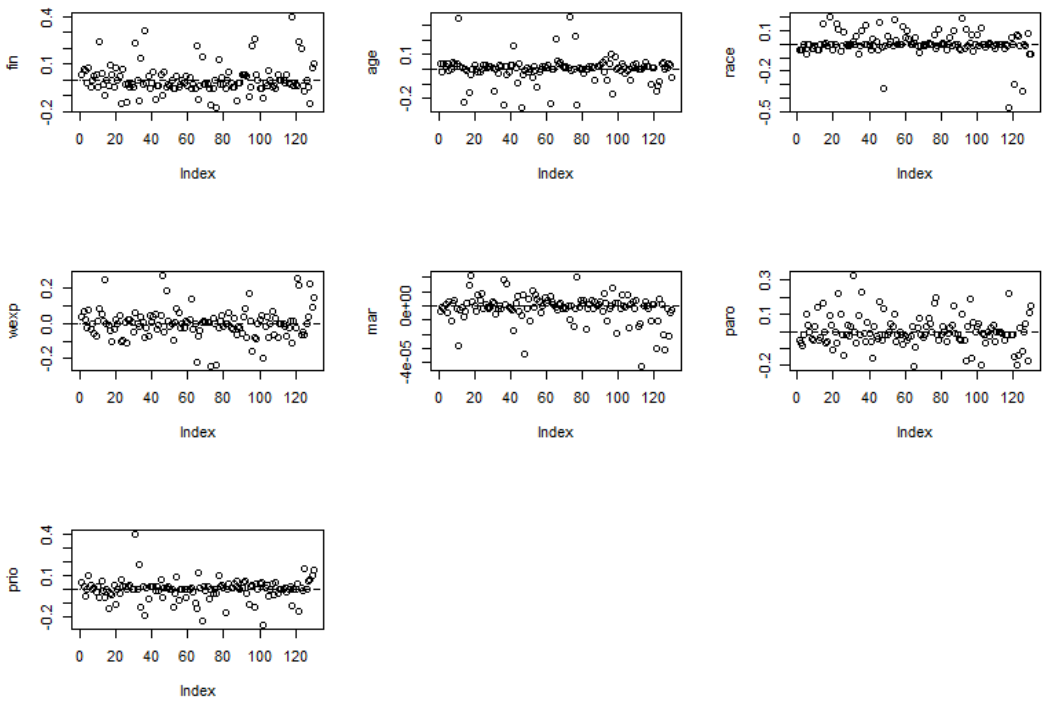


Figura 4.4: Residuos *dfbeta* para covariables de datos de reincidencia. Para cada covariable, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

En la Figura 4.4 se observa que los residuos están centrados con respecto al origen y no presentan patrones definidos. Por lo que, no existe ningún punto influyente en cada una de las covariables del modelo.

4.3. Estudio con datos reales

4.3.1. Conjunto de datos

En este apartado del documento, se lleva a cabo un estudio de supervivencia sobre datos reales de cáncer de mama proporcionados por la Universidad de Yale, en EEUU [1].

- **Situación actual del cáncer de mama**

  El cáncer de mama es un tumor que se origina en las células del seno, es decir, un grupo de células cancerosas que puede crecer e, incluso, invadir los tejidos circundantes o propagarse hacia áreas distintas del cuerpo.

  Actualmente, esta enfermedad es la causa más común de muerte por cáncer en mujeres y, a excepción del cáncer de piel, la primera causa de mortalidad en mujeres de edad superior a los 45 años.

  Según los datos más recientes del Grupo de Trabajo en Estadísticas de Cáncer de Estados Unidos de Norteamérica, en el año 2005, se diagnosticaron 186.467 mujeres y, además, 41.116 mujeres murieron durante ese año por esta enfermedad [13].

  La *American Cancer Society* (ACS) estima que en el año 2008 se diagnosticaron un total de 1,4 millones de casos nuevos de cáncer en Estados Unidos, de los cuales cerca de 250.230 fueron cáncer de mama, lo cual correspondería a un 26 % en las mujeres [14]. A la vez, estima un total de 271.530 mujeres fallecidas por cáncer, de las cuales 40.480 (un 15 %) habrán fallecido por cáncer mamario.

  Según los informes del Ministerio de Sanidad y Consumo en España, en el año 2005, se produjeron 371.934 defunciones. Los tumores se mantuvieron como la segunda causa de muerte en España como responsables del 27 % de las defunciones. En las mujeres, el cáncer de mama fue el más significativo con 5.833 defunciones, con lo cual esta enfermedad tumoral es, hoy por hoy, la primera causa de muerte por cáncer en las mujeres españolas y la sexta causa de muerte global [15].

En este estudio, el conjunto de datos utilizado se compone de 636 pacientes que les ha sido detectado un cáncer de mama. De estos datos se puede destacar que la edad media de los pacientes es 58,1 años y la edad media de muerte es 57,1 años.

Tabla 4.4: Información sobre el conjunto de datos de cáncer de mama. Se incluye el valor mínimo, máximo, media, desviación típica y mediana de la edad de los pacientes en estudio.

Datos	**Edad mínima**	**Edad máxima**	**Media**	**Desv. típica**	**Mediana**
Conjunto completo	24	88	58,11	12,36	58
No supervivientes	24	88	57,12	11,47	57

A pesar de que, los datos recogidos proporcionan información de un periodo de 50 años, el estudio de supervivencia llevado a cabo se realiza a 5, 10 y 15 años. Las variables consideradas en el trabajo se detallan a continuación:

Age_diagnosis Indica la edad de diagnóstico del paciente.

Followup_Time Indica el tiempo, en meses, de seguimiento del paciente.

Censor Es el indicador de censura, codificado como 1 si ha ocurrido el evento (en este caso, la muerte) durante el periodo en que el paciente ha estado en estudio y 0 en caso contrario.

Nuclear_Grade Es el grado del tumor, que representa el "potencial agresivo"del mismo.
Esta variable se cuantifica mediante la siguiente puntuación:
Grado I = 1; Grado II = 2; Grado III = 3.
Siendo el grado III el más agresivo.

Tumor_Size Es el tamaño del tumor (o lesión) en centímetros (cm).

Nodal_Status Indica la existencia de ganglios linfáticos afectados por el tumor.

NPI Es el índice de pronóstico de Nottingham (*Nottingham Prognostic Index*), el cual se utiliza para predecir la supervivencia después de la cirugía para el cáncer de mama.
Se calcula de la siguiente manera: $NPI = [0{,}2 \times S] + N + G$
donde S es el tamaño del tumor, N es el número de ganglios linfáticos afectados, y G es el grado del tumor.
Si el valor obtenido está entre 2,0 y 2,4, la probabilidad de supervivencia a los 5 años es del 93 %; si está entre 2,4 y 3,4, la probabilidad es del 85 %; si está entre 3,4 y 5,4, la probabilidad es del 70 % y si el valor es superior a 5,4, la probabilidad es del 50 % [16].

Her2_Norm Indica el nivel, previamente normalizado, del gen HER2 que posee el paciente en estudio. Se trata de un gen que ayuda al crecimiento de las células normales del cuerpo por medio de la producción de la proteína HER2.
Cuando se tiene un número elevado de copias de este gen, las células (incluyendo las cancerosas) se multiplican más rápidamente. Los expertos piensan que las mujeres con cáncer de mama positivo para HER2 tienen una enfermedad más agresiva, una mayor resistencia a los tratamientos convencionales de quimioterapia y un riesgo mayor de recurrencia que aquellas que no tienen este tipo de cáncer. Además, este gen está altamente relacionado con el gen GRB7, que es un proto-oncogen asociado con el cáncer de mama, tumores de células germinales de testículo, gástricos y de esófago [17].

PR_Norm Muestra el nivel del receptor hormonal de progesterona PR normalizado del

paciente.
La unión de progesterona a una hormona induce un cambio estructural que elimina la acción inhibidora, lo cual significa que PR estimula el crecimiento del tumor [18].

ER_Norm Contiene información sobre el nivel del receptor hormonal de estrógeno normalizado de cada paciente.
La unión de estrógeno al ER estimula la proliferación de células mamarias, con el incremento resultante en la división celular y la replicación del ADN, dando lugar a mutaciones. El metabolismo de los estrógenos produce la interrupción del ciclo celular y la reparación del ADN y, por lo tanto, mitiga o reduce la formación de tumores [19].

auroraB_Norm Indica el nivel normalizado que el paciente posee de la proteína AuroraB.
Niveles anormalmente elevados de AuroraB causan la separación cromosómica desigual durante la división celular, provocando la formación de células con un número anormal de cromosomas, que son a la vez causa de cáncer [20].

stk15_Norm Representa el nivel normalizado de genes STK15 que posee el paciente.
STK15 se considera un potencial de genes de susceptibilidad al cáncer debido a sus funciones en la mitosis celular normal [21].

grb7_Norm Indica el nivel normalizado de proteína GRB7 [22] en el paciente.
Es la proteína unida al receptor de factor de crecimiento 7. Impulsa una forma agresiva de cáncer de mama, por ello, los niveles de proteína GRB7 son un factor importante e independiente en la determinación de un pronóstico para el cáncer de mama [22].

Para poder realizar un estudio completo sobre la supervivencia al cáncer de mama y poder observar la relación que existe entre las distintas variables se elaboran varios experimentos. En el primero de ellos, se incluyen todas las covariables descritas; en el segundo, se excluye la covariable *Age_diagnosis* y en el tercero, se excluye la covariable *NPI*.

4.3.2. Estudio 1: todas las covariables

Estudio de supervivencia a 5 años

1. **Estimación de la función de supervivencia**

   En primer lugar, se estima la función de supervivencia mediante el estimador de Kaplan-Meier, explicado en el apartado 2.4.

La función de supervivencia obtenida de la estimación sobre el conjunto de datos de entrenamiento se incluye más adelante, en la Figura 4.5, junto con la función de supervivencia obtenida mediante el modelo de Cox.

Al igual que en el estudio piloto, la probabilidad de supervivencia al cáncer de mama disminuye con el tiempo.

2. **Selección de variables**

   Para realizar el método de selección de variables, se utilizan las aproximaciones de Efron y de Breslow, obteniendo en ambos casos los mismos resultados. Es por ello que, se incluyen en este trabajo solamente los resultados obtenidos con la aproximación de Efron, ya que es considerada la más exacta.

   En la Tabla 4.5 se muestra el proceso seguido en la selección de variables, donde en cada paso se marca la covariable eliminada del modelo y se indica su p-value correspondiente.

Tabla 4.5: Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 5 años. En cada paso se marca la covariable eliminada del modelo junto con su p-value correspondiente.

Paso	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Age_diagnosis							✓			
Nuclear_Grade				✓						
Tumor_Size								✓		
Nodal_Status		✓								
NPI										
Her2_Norm			✓							
PR_Norm									✓	
ER_Norm					✓					
auroraB_Norm	✓									
stk15_Norm										✓
grb7_Norm						✓				
p-value	0,864	0,793	0,693	0,551	0,469	0,395	0,587	0,151	0,137	0,058

Tras el proceso llevado a cabo se obtiene *NPI* como covariable significativa en el estudio de supervivencia a 5 años, en la que, de manera implícita, influye el tamaño del tumor, el número de ganglios linfáticos afectados y el grado del tumor.

3. **Ajuste del modelo de regresión de Cox**

   Para la estimación de los parámetros del modelo se utiliza el conjunto de datos de entrenamiento, el cual está compuesto por el 70 % de los datos del conjunto inicial. Sin embargo, se eliminan 82 observaciones por valores perdidos.

   El ajuste del modelo de Cox obtenido, considerando únicamente como variable explicativa *NPI*, se muestra en la Tabla 4.6.

Tabla 4.6: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 5 años. Se eliminan 82 observaciones del conjunto de datos por valores perdidos.

n=364, número de eventos=343
(82 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
NPI	0,14839	1,15997	0,04155	3,572	0,000355

Test de razón de verosimilitud=12,57 con 1 grado de libertad, p=0,0003929

Test de Wald=12,76 con 1 grado de libertad, p=0,0003547

Test de puntajes=12,82 con 1 grado de libertad, p=0,0003431

En los test realizados se obtiene un p-value significativamente pequeño (inferior a 0,05), lo cual es evidencia de que, los coeficientes del modelo son significativamente distintos de cero, y por tanto, se puede considerar que el modelo tiene sentido para la variable explicativa considerada.

Además, la covariable *NPI* aparece con un valor positivo para su correspondiente β, lo cual significa que el riesgo de muerte aumenta con la presencia de esta variable, en un factor de 1,16.

La gráfica de la función de supervivencia obtenida mediante el modelo de Cox para *NPI* junto con la obtenida por el estimador de Kaplan-Meier se representan en la Figura 4.5. En esta gráfica se puede observar que la función de supervivencia de Kaplan-Meier es sistemáticamente superior al ajuste del modelo de Cox, pero ambas prácticamente similares.

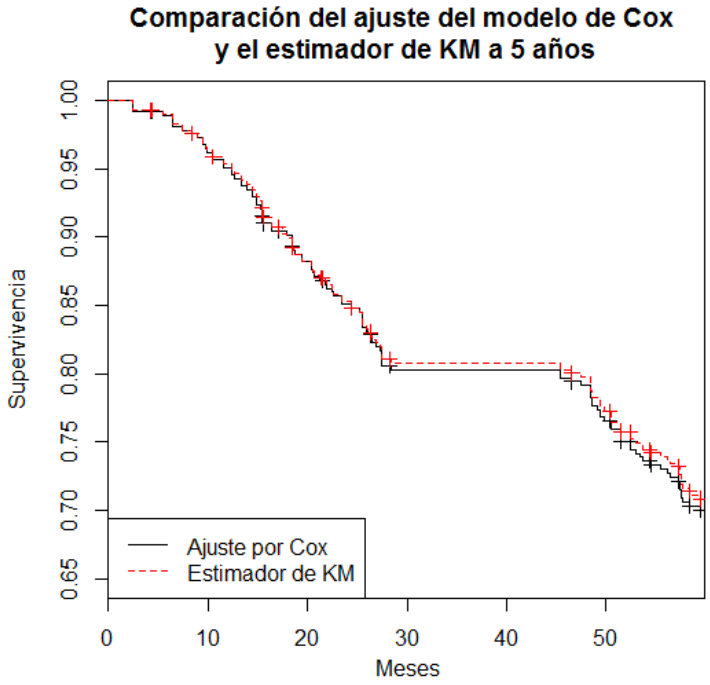


Figura 4.5: Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 5 años con la covariable *NPI*.

4. **Verificación del modelo de Cox**

   Para comprobar que el modelo ajustado de Cox es correcto se utiliza el conjunto de datos de test.

   Para ello, en primer lugar, se realiza el contraste de hipótesis del modelo de Cox, obteniendo como p-value $p = 0{,}558$, de manera que, no existe evidencia significativa al 5 % de que se viole el supuesto del modelo de Cox para la covariable *NPI*.

   En segundo lugar, se realiza el estudio de los residuos tal y como sigue.

   Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

   Mediante los residuos de Cox-Snell se trata de evaluar el ajuste del modelo de Cox. Si el modelo es correcto y la estimación de los β son cercanas a los valores reales, entonces este estimador debería describir aproximadamente una recta de pendiente igual a la unidad.

   Las gráficas de la Figura 4.6 indican que, en general, este modelo se ajusta a los datos a pesar de que en algunos casos la estimación de los β se separe de los valores reales. Los valores negativos de los residuos se corresponden con datos censurados.

   En la primera gráfica se puede observar que, un gran número de datos no se ajusta correctamente a los esperados. Dichos valores se corresponden con pacientes cuya supervivencia es superior a los 5 años. Por este motivo, se representan en la gráfica de la derecha los residuos, calculados de nuevo, sin considerar estos pacientes. En este caso, los valores estimados de β se aproximan a los reales, excepto en la cola de la derecha donde las estimaciones son inestables debido a la censura que presentan los datos.

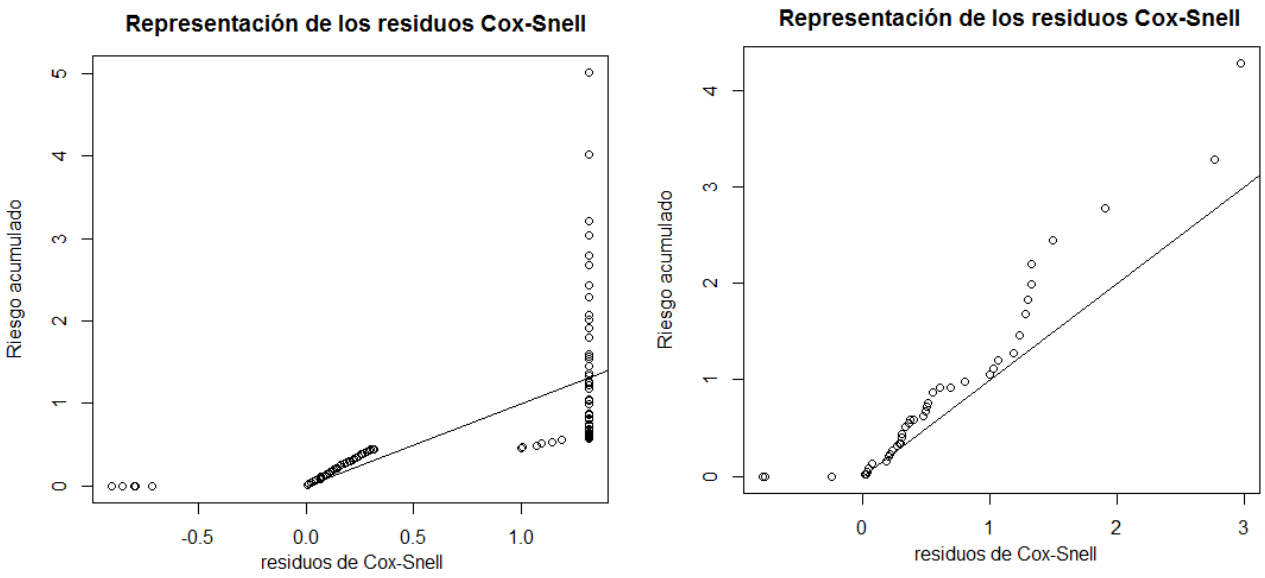


Figura 4.6: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 5 años. Se representan los residuos de Cox-Snell junto con una recta de pendiente 1 para evaluar el ajuste. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 5 años.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

Para determinar la influencia de cada observación en el modelo ajustado, para la covariable *NPI*, se representan en la Figura 4.7 los residuos *dfbeta*. Es decir, para la covariable *NPI*, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

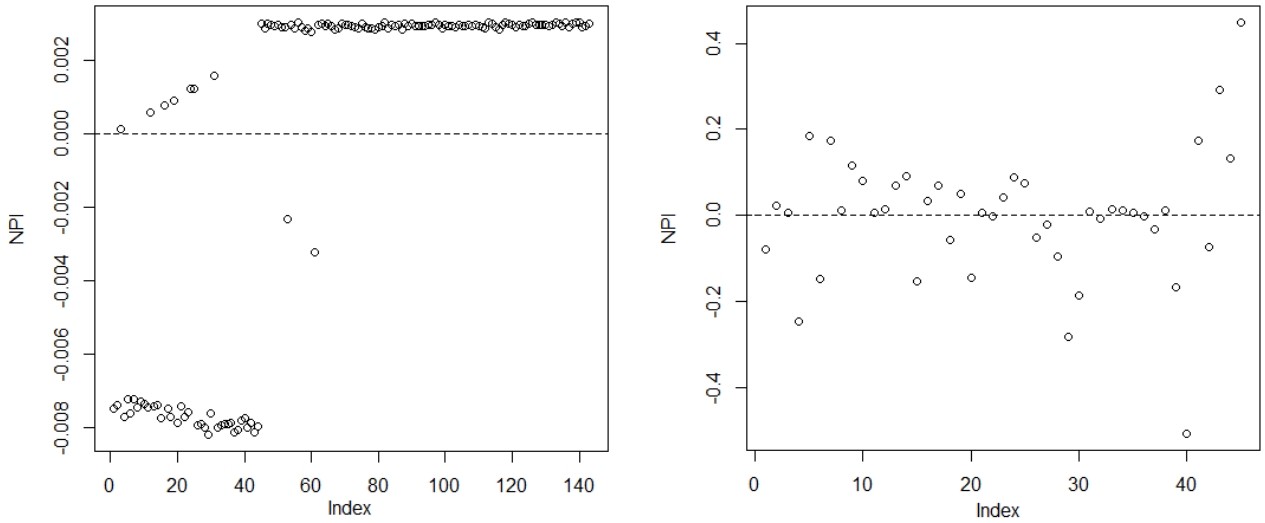


Figura 4.7: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 5 años. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 5 años.

Se observa en la Figura 4.7 que estos residuos se centran en torno al origen, sin dispersarse demasiado.

En la gráfica de la izquierda, en la que se considera el conjunto de test completo, se observan valores constantes de los residuos para la variable *NPI* a partir de un cierto índice. Estos valores se corresponden con datos de pacientes que tienen una supervivencia mayor a 5 años. En la gráfica de la derecha, se eliminan dichos pacientes, observando en este caso que los residuos no presentan ninguna irregularidad. Por lo que, no existe ningún punto influyente en esta covariable.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

Los residuos de *deviance* obtenidos para este estudio se muestran en la Figura 4.8.

En la gráfica de la izquierda se observa un patrón definido que corresponde con pacientes que tienen un tiempo de supervivencia superior a 5 años.

Mientras que, en la gráfica de la derecha, se excluyen estos valores de manera que se obtienen residuos distribuidos en torno al origen. Además, se observa que el número de residuos que está por encima de 1 es superior al que está por debajo de -1, por lo que, hay tendencia a sobrevivir más de lo que predice el modelo.

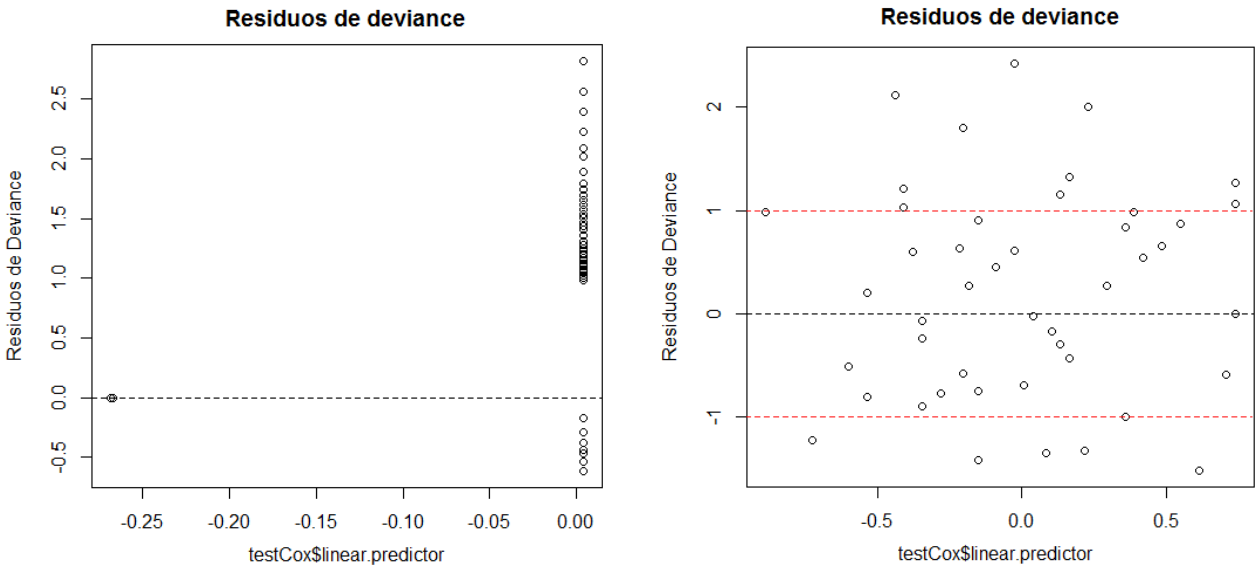


Figura 4.8: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 5 años, para determinar la existencia de *outliers* en el modelo. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 5 años.

Estudio de supervivencia a 10 años

1. **Estimación de la función de supervivencia**

   Al igual que antes, se estima la función de supervivencia a 10 años para los datos de cáncer de mama mediante el estimador de Kaplan-Meier, la cual se representa en la Figura 4.9 junto con la función de supervivencia estimada mediante el modelo de Cox.

2. **Selección de variables**

   Siguiendo la metodología empleada para el estudio de supervivencia a 5 años, en la Tabla 4.7, se incluye el método de selección de variables, señalando la covariable eliminada y su p-value correspondiente.

Tabla 4.7: Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 10 años. En cada paso se marca la covariable eliminada del modelo junto con su p-value correspondiente.

Paso	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age_diagnosis							✓	
Nuclear_Grade	✓							
Tumor_Size								
Nodal_Status								
NPI			✓					
Her2_Norm								
PR_Norm					✓			
ER_Norm				✓				
auroraB_Norm						✓		
stk15_Norm								✓
grb7_Norm		✓						
p-value	0,897	0,828	0,495	0,365	0,237	0,286	0,157	0,126

En la Tabla 4.8 se muestran las covariables significativas en cada paso del método de selección de variables.

Como resultado del proceso, se obtienen como covariables significativas en el estudio de supervivencia a 10 años *Tumor_Size*, *Nodal_Status* y *Her2_Norm*.

Tabla 4.8: Covariables significativas en la selección de variables para datos de cáncer de mama en el estudio de supervivencia a 10 años. En cada paso se marcan las covariables que tienen un p-value inferior a 0,05.

Paso	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Age_diagnosis	✓	✓		✓				
Nuclear_Grade								
Tumor_Size	✓	✓	✓	✓	✓	✓	✓	✓
Nodal_Status	✓	✓	✓	✓	✓	✓	✓	✓
NPI								
Her2_Norm						✓	✓	✓
PR_Norm								
ER_Norm								
auroraB_Norm								
stk15_Norm	✓	✓						
grb7_Norm								

3. **Ajuste del modelo de regresión de Cox**

   En la estimación de los parámetros del modelo se utilizan las covariables significativas determinadas en el apartado anterior. De manera que, se obtiene el modelo de Cox mostrado en la Tabla 4.9. En este caso, el conjunto de datos de entrenamiento se reduce a 358, ya que se eliminan 88 observaciones del conjunto de datos por valores perdidos.

Tabla 4.9: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 10 años. Se eliminan 88 observaciones del conjunto de datos por valores perdidos.

n=358, número de eventos=303
(88 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
Tumor_Size	0,06933	1,07179	0,02484	2,791	0,00525
Nodal_Status	0,50905	1,66371	0,11775	4,323	$1{,}54 \cdot 10^{-05}$
Her2_Norm	0,15549	1,16823	0,06198	2,509	0,01211

Test de razón de verosimilitud=33 con 3 grados de libertad, p=$3{,}214 \cdot 10^{-07}$

Test de Wald=35,58 con 3 grados de libertad, p=$9{,}199 \cdot 10^{-08}$

Test de puntajes=36,26 con 3 grados de libertad, p=$6{,}6 \cdot 10^{-08}$

Se obtiene evidencia de que, los coeficientes del modelo son significativamente distintos de cero, y por tanto, se considera que el modelo tiene sentido para las covariables consideradas.

Todas las variables en el modelo ajustado aparecen con un valor positivo para su correspondiente β, lo cual significa que el riesgo de muerte aumenta con la presencia de las mismas.

En la Figura 4.9 se representa la función de supervivencia obtenida mediante el estimador de Kaplan-Meier y la obtenida por el modelo de Cox teniendo en cuenta las tres covariables seleccionadas. En dicha gráfica se observa que el ajuste del modelo de Cox es prácticamente similar a la función de supervivencia de Kaplan-Meier.

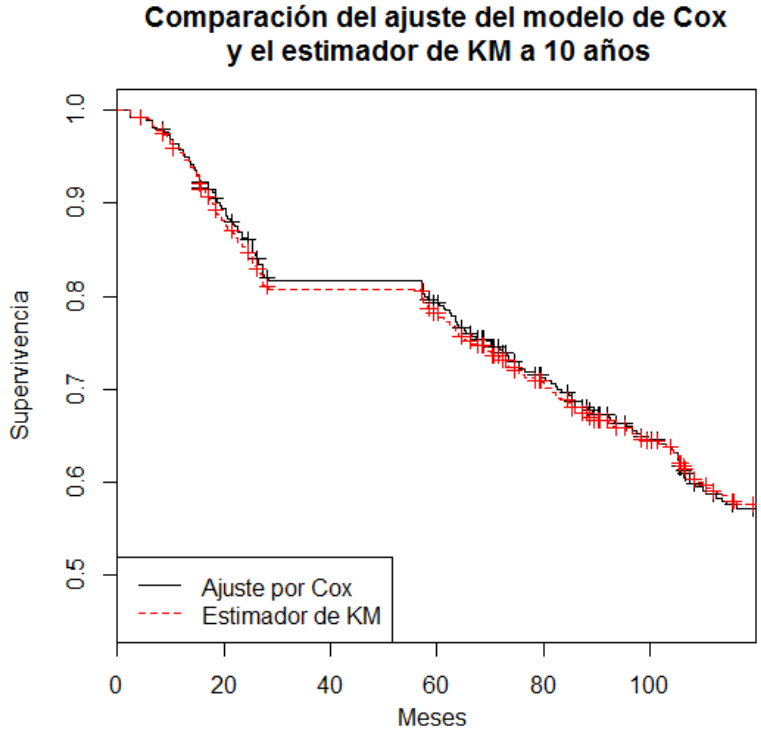


Figura 4.9: Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 10 años con las variables *Tumor_Size*, *Nodal_Status* y *Her2_Norm*.

4. **Verificación del modelo de Cox**

   En primer lugar, se realiza el contraste de hipótesis del modelo de Cox, obteniendo la información mostrada en la Tabla 4.10.

Tabla 4.10: Resultado de contraste del modelo de Cox para cáncer de mama en el estudio de supervivencia a 10 años, considerando como hipótesis nula el cumplimiento del supuesto de Cox. Para cada covariable se incluye el p-value resultante.

	p
Tumor_Size	0,3262
Nodal_Status	0,3482
Her2_Norm	0,0642
GLOBAL	0,1131

Puesto que los valores obtenidos para cada p-value son superiores a 0,05, no existe evidencia significativa al 5 % de que se viole el supuesto del modelo de Cox, ni desde el punto de vista global, ni para cada covariable.

Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

En las gráficas de la Figura 4.10 se observa que, en general, el modelo se aproxima a los valores reales. Además, este modelo se ajusta bastante mejor a los datos que el obtenido para la supervivencia a 5 años (Figura 4.6). Por tanto, el modelo planteado para supervivencia a 10 años es más fiable. En la gráfica de la derecha, se representan los residuos de Cox-Snell para datos sin considerar tiempos de supervivencia superiores a 10 años, donde se obtiene un mejor ajuste.

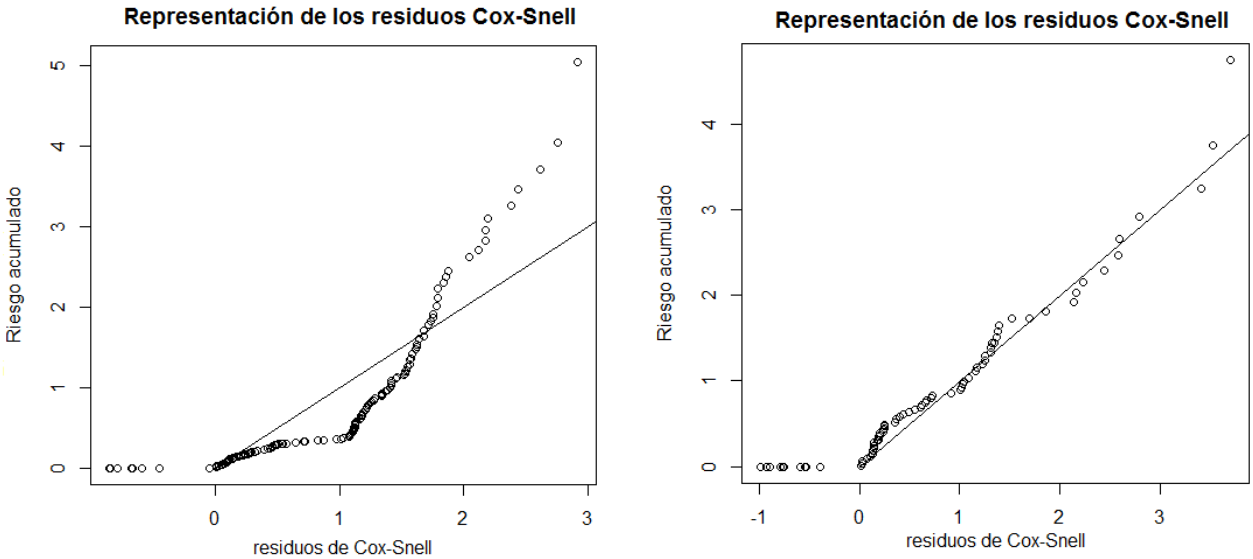
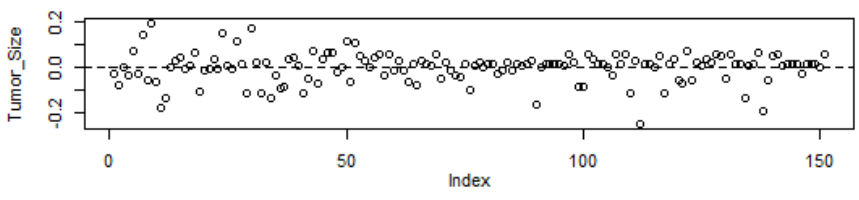


Figura 4.10: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 10 años. Se representan los residuos de Cox-Snell con una recta de pendiente 1 para evaluar el ajuste. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes con supervivencia superior a 10 años.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

En la Figura 4.11 se representan los residuos *dfbeta* para las covariables *Tumor_Size*, *Nodal_Status* y *Her2_Norm*, observando que, no existe ningún punto influyente, ya que estos residuos se centran en torno al origen y no presentan ninguna irregularidad en sus gráficas, salvo algún valor anómalo en la covariable *Her2_Norm*.



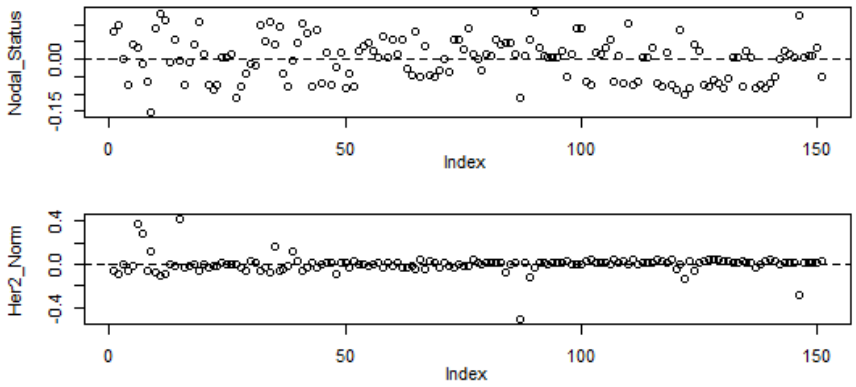


Figura 4.11: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 10 años. Para cada covariable, se representa la observación por el cambio de escala aproximada del coeficiente después de la eliminación de la observación del modelo.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

En la Figura 4.12 se muestran los residuos de *deviance* obtenidos en este estudio. En la gráfica de la izquierda se observa un patrón extraño en los residuos, que se corresponde con datos de pacientes que tienen una supervivencia superior a los 10 años. Por ello, se representa en la gráfica de la derecha el valor de este tipo de residuos sin considerar datos con supervivencia superior a 10 años. En este caso, los residuos no siguen ningún patrón definido ni tampoco se alejan demasiado del origen y, al igual que en el estudio de supervivencia a 5 años, se observa que, el modelo predice tiempos de supervivencia inferiores a los reales.

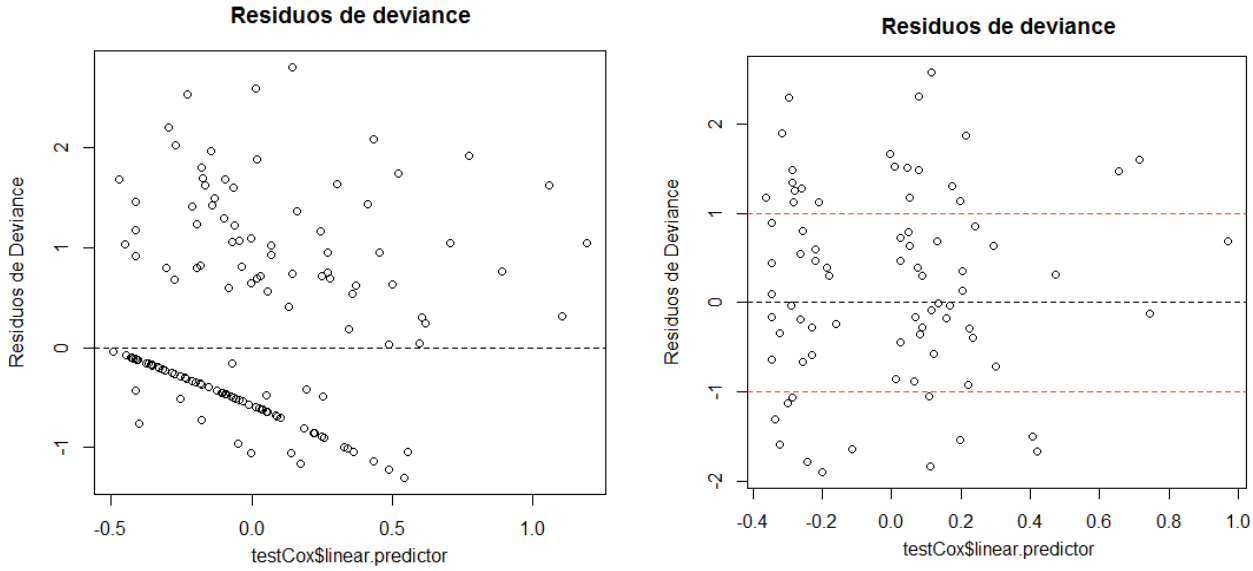


Figura 4.12: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 10 años, para determinar la existencia de *outliers* en el modelo. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 10 años.

Estudio de supervivencia a 15 años

1. **Estimación de la función de supervivencia**

   Al igual que en los casos anteriores, se estima la función de supervivencia a 15 años para los datos de cáncer de mama, la cual se representa junto con la obtenida mediante el modelo de Cox en la Figura 4.13.

2. **Selección de variables**

   En la Tabla 4.11 se incluye el método de selección de variables, señalando la covariable eliminada y su p-value correspondiente y, en la Tabla 4.12, se muestran las covariables significativas en cada paso.

Tabla 4.11: Selección de variables para datos de cáncer de mama en el estudio de supervivencia a 15 años. En cada paso se marca la covariable eliminada del modelo junto con su p-value correspondiente.

Paso	**1**	**2**	**3**	**4**	**5**	**6**
Age_diagnosis						
Nuclear_Grade						✓
Tumor_Size				✓		
Nodal_Status	✓					
NPI						
Her2_Norm						
PR_Norm			✓			
ER_Norm						
auroraB_Norm					✓	
stk15_Norm						
grb7_Norm		✓				
p-value	0,824	0,760	0,298	0,206	0,183	0,056

Las covariables significativas obtenidas para el estudio de supervivencia a 15 años son *Age_diagnosis*, *NPI*, *Her2_Norm*, *ER_Norm* y *stk15_Norm*.

Tabla 4.12: Covariables significativas en la selección de variables para datos de cáncer de mama en el estudio de supervivencia a 15 años. En cada paso se marcan las covariables que tienen un p-value inferior a 0,05.

Paso	**1**	**2**	**3**	**4**	**5**	**6**
Age_diagnosis	✓	✓	✓	✓	✓	✓
Nuclear_Grade					✓	
Tumor_Size						
Nodal_Status						
NPI		✓	✓	✓	✓	✓
Her2_Norm			✓	✓	✓	✓
PR_Norm						
ER_Norm	✓	✓	✓	✓	✓	✓
auroraB_Norm						
stk15_Norm	✓	✓	✓	✓	✓	✓
grb7_Norm						

3. **Ajuste del modelo de regresión de Cox**

   El modelo de Cox ajustado para las covariables significativas se muestra en la Tabla 4.13, donde se obtiene evidencia de que los coeficientes del modelo son significativamente distintos de cero, por tanto, se considera que el modelo tiene sentido para las covariables consideradas. En este caso, se eliminan del conjunto de datos 145 observaciones por valores perdidos.

Tabla 4.13: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años. Se eliminan 145 observaciones del conjunto de datos por valores perdidos.

n=301, número de eventos=246
(145 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
Age_diagnosis	0,011881	1,011952	0,005688	2,089	0,0367
NPI	0,249293	1,283119	0,049150	5,072	$3{,}93 \cdot 10^{-07}$
Her2_Norm	0,147851	1,159340	0,060716	2,435	0,0149
ER_Norm	−0,155137	0,856298	0,071584	−2,167	0,0302
stk15_Norm	0,169612	1,184845	0,078281	2,167	0,0303

Test de razón de verosimilitud=45,06 con 5 grados de libertad, p=$1{,}412 \cdot 10^{-08}$

Test de Wald=46,86 con 5 grados de libertad, p=$6{,}071 \cdot 10^{-09}$

Test de puntajes=47,88 con 5 grados de libertad, p=$3{,}756 \cdot 10^{-09}$

Todas las covariables en el modelo ajustado aparecen con un valor positivo para su correspondiente β, excepto la covariable *ER_Norm* que tiene valor negativo. Esto significa que, cuanto mayor sea el valor de las covariables *Age_diagnosis*, *NPI*, *Her2_Norm* y *stk15_Norm* mayor riesgo de muerte existirá, mientras que *ER_Norm*, por el contrario, disminuye el riesgo de muerte, es decir, cuanto mayor sea el nivel de estrógeno menos riesgo de muerte tendrá el paciente.

En la Figura 4.13 se representa la función de supervivencia obtenida mediante el estimador de Kaplan-Meier junto con la obtenida por el modelo de Cox teniendo en cuenta las covariables seleccionadas como significativas. En esta gráfica se observa que, el ajuste del modelo de Cox es sistemáticamente inferior a la función de supervivencia de Kaplan-Meier, pero ambas bastante similares.

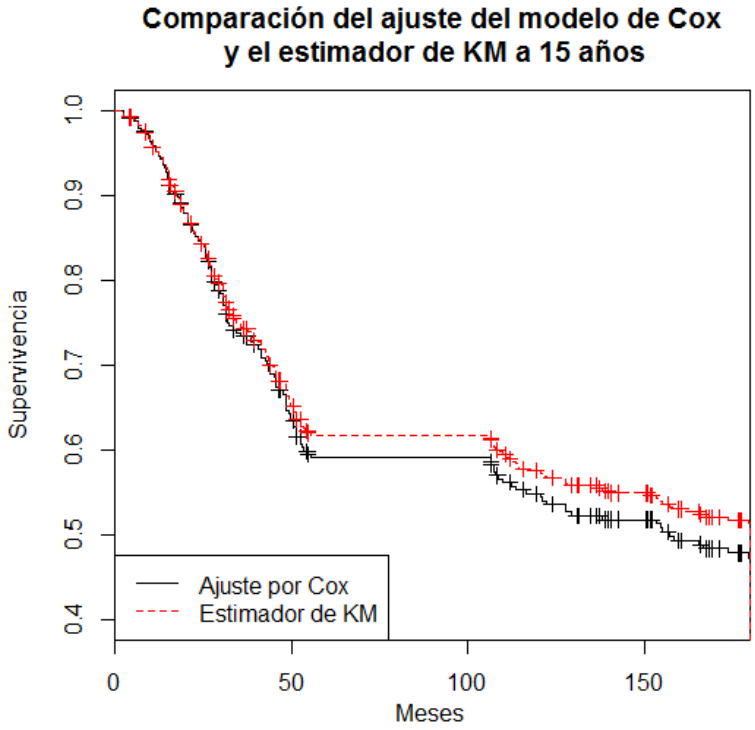


Figura 4.13: Comparación del ajuste del modelo de Cox y el estimador de Kaplan-Meier para datos de cáncer de mama en el estudio de supervivencia a 15 años con las variables *Age_diagnosis*, *NPI*, *Her2_Norm*, *ER_Norm* y *stk15_Norm*.

4. **Verificación del modelo de Cox**

   En primer lugar, se realiza el contraste de hipótesis del modelo de Cox, obteniendo la información mostrada en la Tabla 4.14.

   De manera que, no existe evidencia significativa al 5 % de que se viole el supuesto del modelo de Cox, ni desde el punto de vista global, ni para cada covariable.

Tabla 4.14: Resultado de contraste del modelo de Cox para cáncer de mama en el estudio de supervivencia a 15 años, considerando como hipótesis nula el cumplimiento del supuesto de Cox. Para cada covariable se incluye el p-value resultante.

	p
Age_diagnosis	0,8841
NPI	0,0997
Her2_Norm	0,6455
ER_Norm	0,7696
stk15_Norm	0,3070
GLOBAL	0,7273

Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

La Figura 4.14 indica que, en general, este modelo se ajusta bien a los datos. En la gráfica de la izquierda, en la cual se considera todo el conjunto de test, se observa mayor dispersión entre la estimación y los valores reales de los β. Sin embargo, en la gráfica de la derecha, se excluyen los pacientes que tienen un tiempo de supervivencia superior a 15 años, obteniendo un mejor ajuste a los datos.

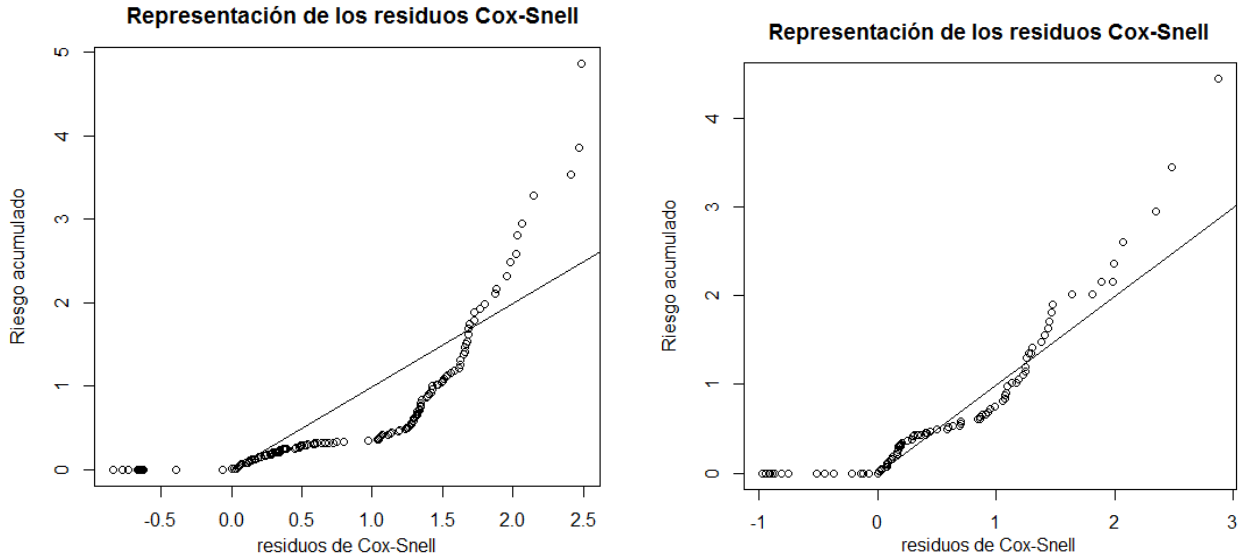


Figura 4.14: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años. Se representan los residuos de Cox-Snell junto con una recta de pendiente 1 para evaluar el ajuste. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 15 años.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

En la Figura 4.15 se muestran los residuos *dfbeta* para cada covariable seleccionada, donde se observa que éstos se concentran en torno al origen y no presentan ninguna irregularidad.

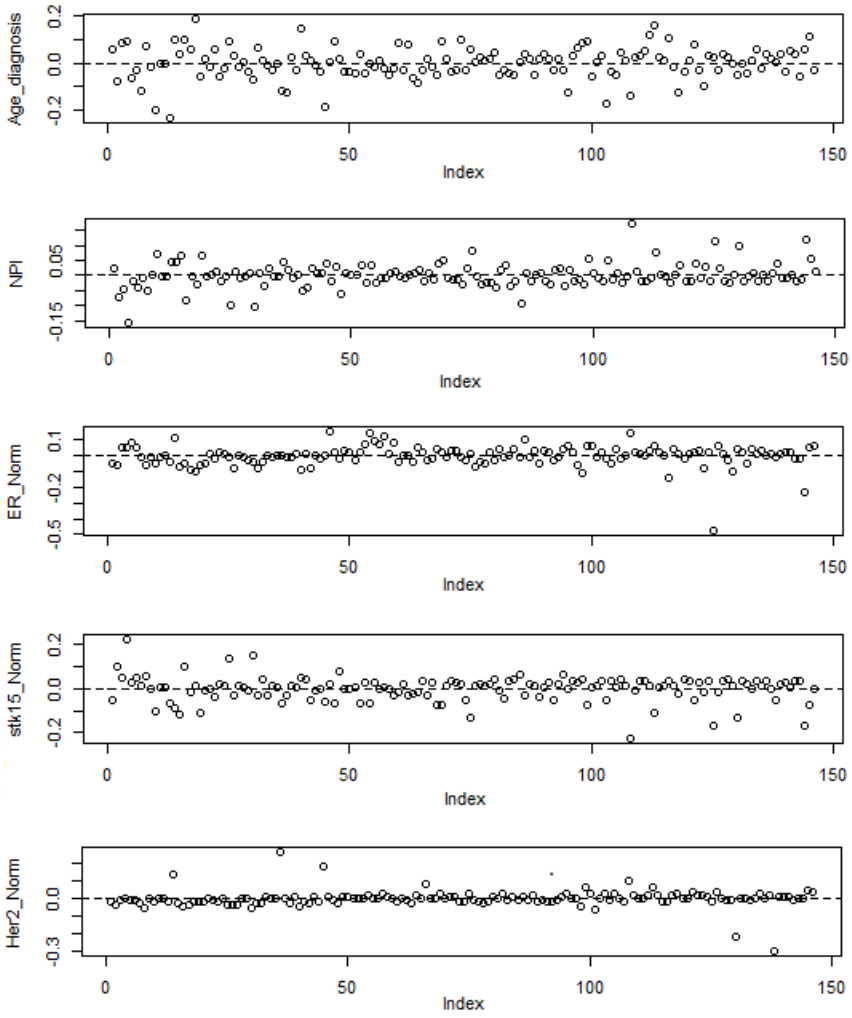


Figura 4.15: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años. Para cada covariable, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

En la Figura 4.16 se muestran los residuos de *deviance*. En la gráfica de la izquierda, se observa un patrón definido, que se corresponde con pacientes que tienen un tiempo de supervivencia superior a los 15 años. En la gráfica de la derecha, se eliminan estos datos, y en este caso, se obtienen residuos que no siguen ningún patrón definido ni tampoco están alejados del origen. Se puede observar que hay una tendencia a sobrevivir más de lo que predice el modelo, ya que el número de residuos que está por encima de 1 es claramente superior al que está por debajo de -1.

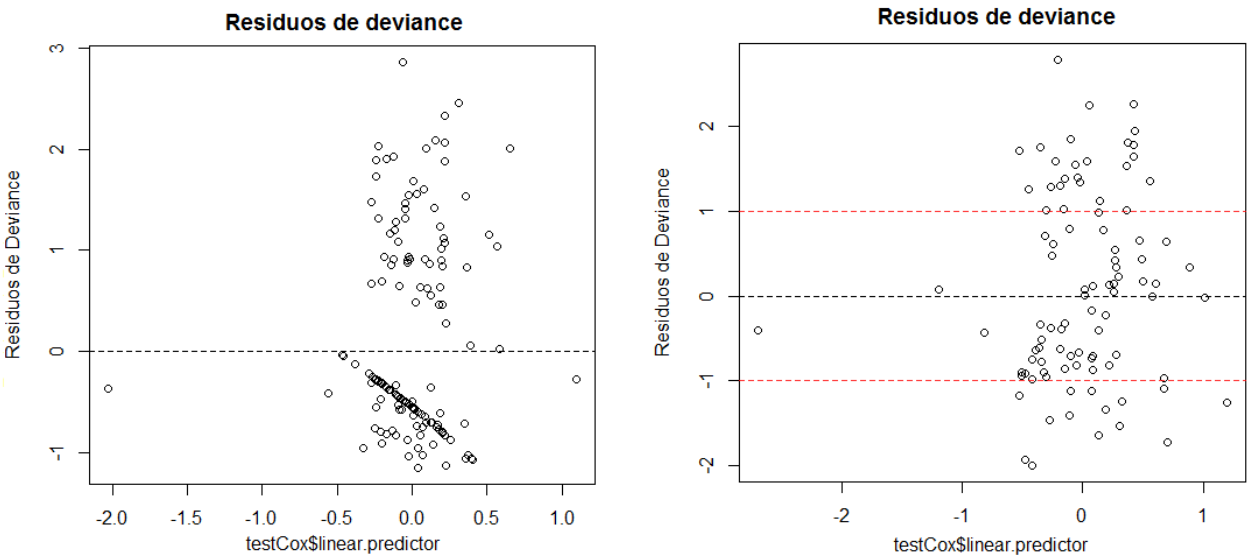


Figura 4.16: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años, para determinar la existencia de *outliers* en el modelo. En la gráfica de la izquierda se representan los residuos del conjunto de test completo. En la gráfica de la derecha se excluyen los pacientes cuya supervivencia es superior a 15 años.

Conclusiones

En este primer estudio realizado en el que se consideran todas las covariables existentes en los datos, se obtiene como factor pronóstico de supervivencia a 5 años *NPI*, variable en la cual influye el tamaño del tumor, el número de ganglios linfáticos afectados y el grado del tumor. Para la supervivencia a 10 años, se obtienen *Tumor_Size*, *Nodal_Status* y *Her2_Norm*, lo cual guarda relación con la supervivencia a 5 años. Para la supervivencia a 15 años, dado que, son bastantes años para estudiar la supervivencia en este tipo de cáncer, influye *Age_diagnosis* y *stk15_Norm* aumentando el riesgo de muerte y *ER_Norm* disminuyendo el riesgo [14]. Además de los factores que ya aparecen en los estudios de supervivencia a 5 y 10 años, es decir, *NPI* y *Her2_Norm*.

De esta manera, en el estudio de supervivencia a 5 años ya se obtienen parte de las covariables, de hecho las más influyentes, que serán un factor pronóstico para la supervivencia a 10 años y 15 años.

En cuanto a la probabilidad de supervivencia, se obtiene una supervivencia global a 5 años del 70 % frente a un 62 % a 10 años y un 53 % a 15 años, es decir, disminuye la supervivencia a medida que aumentan los años que se padece la enfermedad.

En general, con respecto al estudio de residuos, se puede concluir que el modelo estimado es adecuado y existe una tendencia a sobrevivir más de lo que se predice.

4.3.3. Estudio 2: sin *Age_diagnosis*

En este experimento, se excluye la covariable *Age_diagnosis* para poder determinar cómo influye la edad con respecto al resto de variables.

Puesto que, la función de supervivencia ya ha sido estimada en el experimento anterior y el hecho de eliminar una sola covariable del estudio no afecta apenas nada a dicha función, no se realiza la estimación de la función de supervivencia mediante el estimador de Kaplan-Meier.

En los estudios a 5 y 10 años, en el método de selección de variables se obtienen las mismas covariables significativas que en el estudio llevado a cabo en el apartado anterior, puesto que, la covariable *Age_diagnosis* suprimida, no influye en las mismas. De modo que, los pasos realizados en el estudio son similiares y es por ello que no se incluyen en este apartado.

Estudio de supervivencia a 15 años

1. **Selección de variables**

   En el análisis anterior, para la supervivencia a 15 años se obtienen como covariables significativas *Age_diagnosis*, *NPI*, *Her2_Norm*, *ER_Norm* y *stk15_Norm*. Puesto que, en este caso se excluye *Age_diagnosis*, tras el método de selección de variables llevado a cabo, las covariables significativas para el estudio de supervivencia a 15 años se reducen, siendo ahora *NPI*, *Her2_Norm* y *stk15_Norm*.

2. **Ajuste del modelo de regresión de Cox**

   El modelo de Cox ajustado para las covariables significativas obtenidas se muestra en la Tabla 4.15, donde se obtiene evidencia de que los coeficientes del modelo son significativamente distintos de cero. Por tanto, se considera que este modelo es correcto para las covariables consideradas. Se eliminan 139 observaciones del conjunto de datos por valores perdidos.

Tabla 4.15: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. Se eliminan 139 observaciones por valores perdidos.

n=307, número de eventos=251
(139 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
NPI	0,22643	1,25411	0,04698	4,820	$1{,}44 \cdot 10^{-06}$
Her2_Norm	0,16023	1,17378	0,05953	2,692	0,00711
stk15_Norm	0,15453	1,16711	0,07560	2,044	0,04095

Test de razón de verosimilitud=38,2 con 3 grados de libertad, p=$2{,}568 \cdot 10^{-08}$

Test de Wald=40,08 con 3 grados de libertad, p=$1{,}023 \cdot 10^{-08}$

Test de puntajes=40,92 con 3 grados de libertad, p=$6{,}801 \cdot 10^{-09}$

Todas las covariables en el modelo ajustado aparecen con un valor positivo para su correspondiente β, por lo que, cuanto mayor sea el valor de las covariables mayor riesgo de muerte existirá.

En este caso, la función de supervivencia obtenida mediante el estimador de Kaplan-Meier y la obtenida por el modelo de Cox es similar a la del estudio anterior. Por este motivo, no se incluye la gráfica en este apartado.

3. **Verificación del modelo de Cox**

Siguiendo el mismo procedimiento que en el caso anterior, se realiza el contraste de hipótesis del modelo de Cox donde se obtiene para *NPI* un p-value $p = 0{,}108$, para *Her2_Norm* $p = 0{,}666$ y para *stk15_Norm* $p = 0{,}341$. Los valores obtenidos son muy similares a los del estudio anterior para las mismas covariables (Tabla 4.14) y, por tanto, no existe evidencia significativa al 5 % de que se viole el supuesto del modelo de Cox.

Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

En la Figura 4.17 se representan los residuos de Cox-Snell. Al igual que en los estudios anteriores, este modelo se ajusta bien a los datos. Aunque se debe tener en cuenta que, en el cálculo de residuos, se excluyen del conjunto de datos pacientes con tiempos de supervivencia superiores a los 15 años, mejorando así el ajuste.

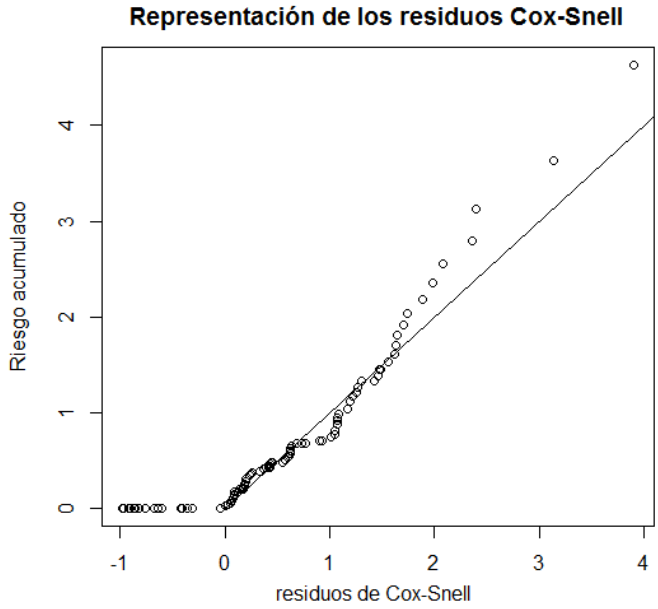


Figura 4.17: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis* y sin pacientes con tiempos de supervivencia superiores a 15 años. Se representan los residuos de Cox-Snell junto con una recta de pendiente 1 para evaluar el ajuste.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

La Figura 4.18 representa los residuos *dfbeta* para cada covariable del modelo. No presentan ningún punto influyente, ya que se concentran en un rango pequeño en torno al origen.

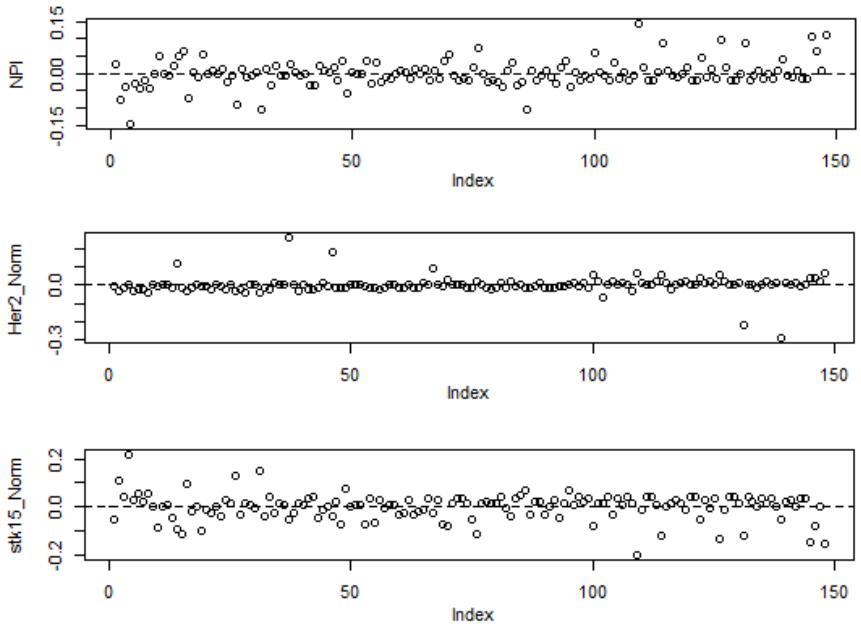


Figura 4.18: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*. Para cada covariable, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

Los residuos de *deviance* obtenidos presentan un patrón definido, el cual se corresponde, tal y como se ha explicado en los casos anteriores, con pacientes que tienen un tiempo de supervivencia superior a los 15 años.

Por ello, en la Figura 4.19 se representan estos residuos sin considerar dichos pacientes, donde se puede observar que no se alejan demasiado del origen.

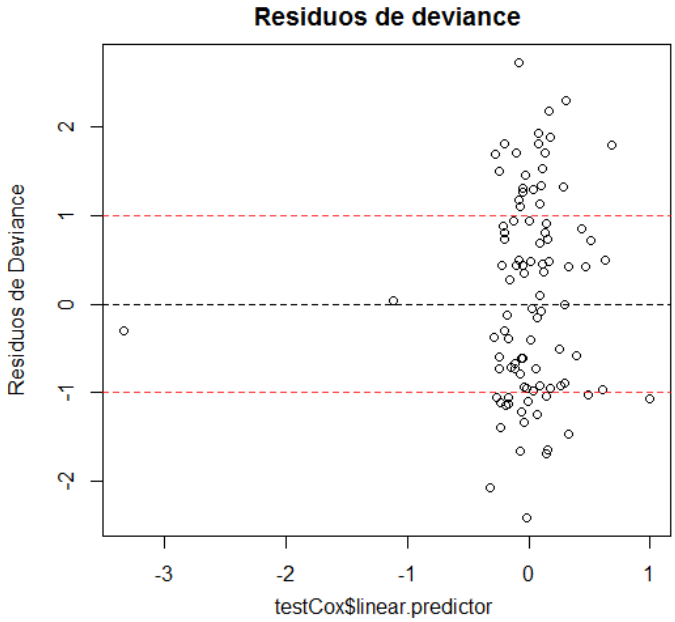


Figura 4.19: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *Age_diagnosis*, para determinar la existencia de *outliers* en el modelo.

Conclusiones

En el análisis realizado se observa que al eliminar del estudio la covariable *Age_diagnosis*, en la supervivencia a 15 años, cambian las variables significativas.

Ahora, éstas son *NPI*, *Her2_Norm* y *stk15_Norm*, entre las cuales desparece *ER_Norm*. Lo cual es evidencia de que, está altamente relacionada con la edad, en concreto, el número de estrógenos disminuye considerablemente con la edad en las mujeres [23].

4.3.4. Estudio 3: sin *NPI*

El objetivo de este estudio es determinar la relación que tiene la covariable *NPI* con el resto.

En el caso del estudio de supervivencia a 10 años, no se obtiene esta covariable como significativa, de manera que, no cambiará nada con respecto al primer estudio realizado. Por esta razón, no se incluye esta parte en este tercer experimento sin *NPI*.

Además, la función de supervivencia estimada mediante Kaplan-Meier es la misma que en los estudios anteriores y, por ello, tampoco se incluye en este apartado.

Estudio de supervivencia a 5 años

1. **Selección de variables**

   En el estudio de supervivencia a 5 años realizado en el primer experimento, se obtiene como covariable significativa *NPI*. En este caso, al suprimir esta variable, aparecen *Tumor_Size* y

Nodal_Status como significativas, ya que éstas forman parte implícitamente de la covariable eliminada.

2. **Ajuste del modelo de regresión de Cox**

   En la Tabla 4.16, se muestra el modelo de Cox ajustado para las covariables significativas obtenidas y se tiene evidencia de que el modelo es correcto para estas variables. Además, se eliminan 62 observaciones del conjunto de datos por valores perdidos.

Tabla 4.16: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*. Se eliminan 62 observaciones del conjunto por valores perdidos.

n=384, número de eventos=362
(62 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
Tumor_Size	0,05494	1,05648	0,02296	2,393	0,0167
Nodal_Status	0,25999	1,29692	0,10699	2,430	0,0151

Test de razón de verosimilitud=13,4 con 2 grados de libertad, p=0,001232

Test de Wald=14,17 con 2 grados de libertad, p=0,0008391

Test de puntajes=14,25 con 2 grados de libertad, p=0,0008053

Los coeficientes β de las variables seleccionadas aparacen con un valor positivo, lo que significa que, aumentan el riesgo de muerte.

3. **Verificación del modelo de Cox**

   Siguiendo con la metodología llevada a cabo en los estudios anteriores, se realiza el contraste de hipótesis del modelo de Cox con los datos de test, obteniendo un valor de p-value superior a 0,05 para la covariable *Tumor_Size* ($p = 0{,}1368$). Mientras que, por el contrario, para la covariable *Nodal_Status* se obtiene un p-value $p = 0{,}0386$. Por lo que, en este caso, el modelo de regresión no es correcto para la segunda covariable.

   Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

   Tras el cálculo de residuos Cox-Snell, se tiene un ajuste que falla en valores grandes de los residuos debido a que, tal y como se ha comentado anteriormente, se incluyen datos de pacientes con tiempos de supervivencia superiores a los 5 años y, en este caso, el supuesto del modelo de Cox se viola para una de las covariables.

   Excluyendo del conjunto de datos los pacientes cuyo tiempo de supervivencia es superior a los 5 años, Figura 4.20, se muestra que, en general, este modelo se ajusta a los datos, excepto en la cola de la derecha donde las estimaciones son inestables debido a la censura de los datos.

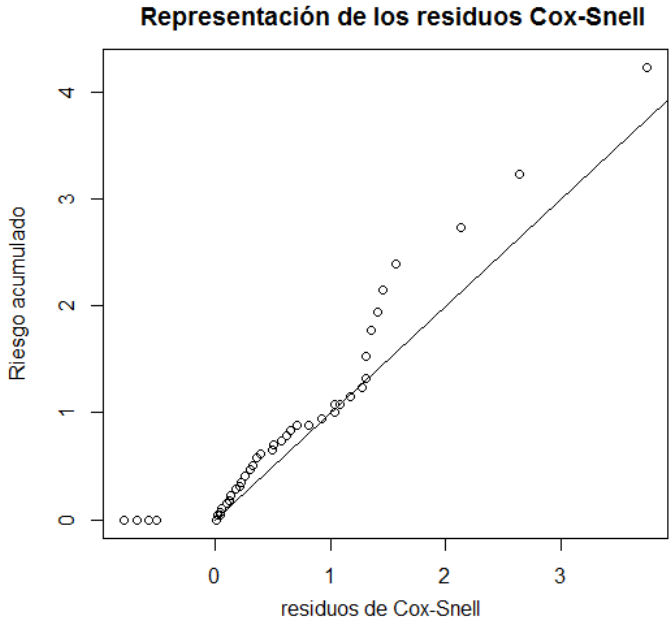


Figura 4.20: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI* y sin pacientes con tiempos de supervivencia superiores a 5 años. Se representan los residuos de Cox-Snell junto con una recta de pendiente 1 para evaluar el ajuste.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

En la Figura 4.21 se representan los residuos *dfbeta* para cada covariable significativa del modelo y se observa que éstos no presentan ninguna irregularidad para *Tumor_Size*. Sin embargo, para *Nodal_Status* se observan ciertas irregularidades, debido a que ésta viola el supuesto del modelo de Cox.

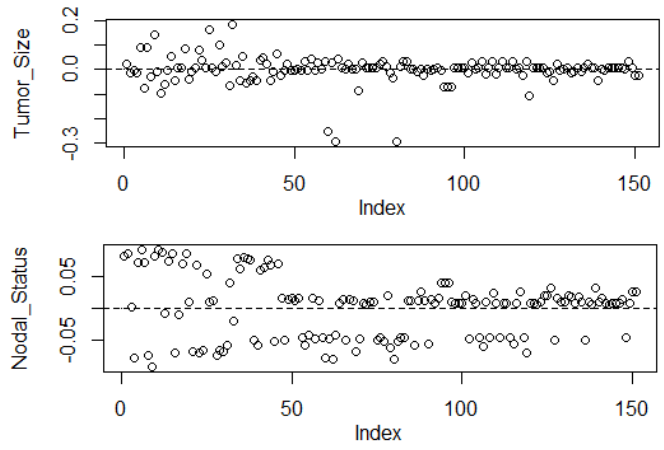


Figura 4.21: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI* y sin pacientes con tiempos de supervivencia superiores a 5 años. Para cada covariable, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

Los residuos de *deviance* obtenidos siguen un patrón definido, el cual se corresponde con tiempos de supervivencia superiores a los 5 años. Mientras que el resto de valores indican que no existen valores atípicos en el modelo.

Por tanto, en la Figura 4.22, se representan estos residuos sin considerar pacientes con tiempos de supervivencia superiores a 5 años, donde se observa, al igual que en los estudios anteriores, que hay tendencia a sobrevivir más de lo que predice el modelo.

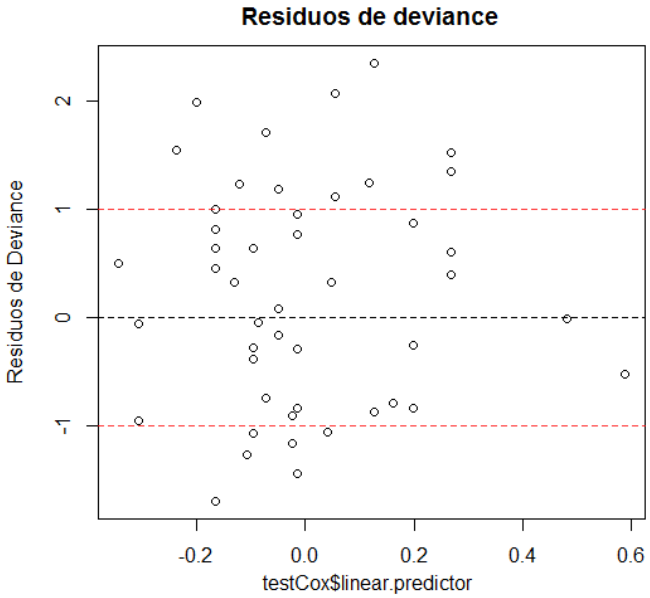


Figura 4.22: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 5 años sin *NPI*, para determinar la existencia de *outliers* en el modelo.

Estudio de supervivencia a 15 años

1. **Selección de variables**

   En el estudio de supervivencia a 15 años realizado en el primer experimento, se obtienen cinco covariables significativas entre las cuales se incluye *NPI*. Al suprimir esta variable del estudio, tras el proceso de selección de variables, aparecen *Tumor_Size* y *Nodal_Status* en lugar de *NPI*, al igual que ocurre en el experimento para supervivencia a 5 años.

2. **Ajuste del modelo de regresión de Cox**

   En la Tabla 4.17 se muestran los parámetros obtenidos del ajuste del modelo de Cox. Se obtiene evidencia de que los coeficientes del modelo son significativamente distintos de cero, y por tanto, se considera que el modelo tiene sentido para las covariables consideradas.

   Todas las variables en este modelo ajustado aparecen con un valor positivo para su correspondiente β, excepto *ER_Norm*. Por tanto, cuanto mayor sea el valor de las covariables

Age_diagnosis, *Tumor_Size*, *Nodal_Status*, *Her2_Norm* y *stk15_Norm* aumentará el riesgo de muerte. Mientras que para la covariable *ER_Norm*, por el contrario, disminuye el riesgo de muerte, es decir, cuanto mayor sea el nivel de estrógeno menor será el riesgo de muerte.

Tabla 4.17: Parámetros del modelo de Cox para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. Se eliminan 128 observaciones por valores perdidos.

n=318, número de eventos=260
(128 observaciones eliminadas por valores perdidos)

	coef	*exp(coef)*	*se(coef)*	*z*	$Pr(>\|z\|)$
Age_diagnosis	0,012302	1,012378	0,005656	2,175	0,02963
Tumor_Size	0,090140	1,094328	0,027895	3,231	0,00123
Nodal_Status	0,517715	1,678189	0,130334	3,972	$7{,}12 \cdot 10^{-05}$
Her2_Norm	0,171026	1,186522	0,062076	2,755	0,00587
ER_Norm	−0,144921	0,865090	0,072652	−1,995	0,04607
stk15_Norm	0,179098	1,196137	0,076706	2,335	0,01955

Test de razón de verosimilitud=50,9 con 6 grados de libertad, p=$3{,}098 \cdot 10^{-09}$

Test de Wald=53,49 con 6 grados de libertad, p=$9{,}329 \cdot 10^{-10}$

Test de puntajes=55,24 con 6 grados de libertad, p=$4{,}136 \cdot 10^{-10}$

3. **Verificación del modelo de Cox**

Como resultado del contraste de hipótesis del modelo de Cox con los datos de test se obtiene que, no existe evidencia significativa al 5 % de que se viole el supuesto del modelo de Cox, ni desde el punto de vista global, ni para cada covariable. Ya que para *Age_diagnosis* se obtiene un p-value $p = 0{,}784$, para *Tumor_Size* $p = 0{,}590$, para *Nodal_Status* $p = 0{,}341$, para *Her2_Norm* $p = 0{,}871$, para *ER_Norm* $p = 0{,}659$ y para *stk15_Norm* $p = 0{,}896$; y desde el punto de vista global se tiene un p-value $p = 0{,}944$.

El estudio de residuos llevado a cabo se muestra a continuación.

Comprobación de la hipótesis global del modelo: Residuos de Cox-Snell

En la Figura 4.23, se muestra que este modelo se ajusta bien a los datos, salvo en pacientes que presentan un tiempo de supervivencia superior a 15 años que, tal y como se ha explicado anteriormente, estos valores no se ajustan correctamente y, por tanto, no se incluyen en el cálculo de estos residuos.

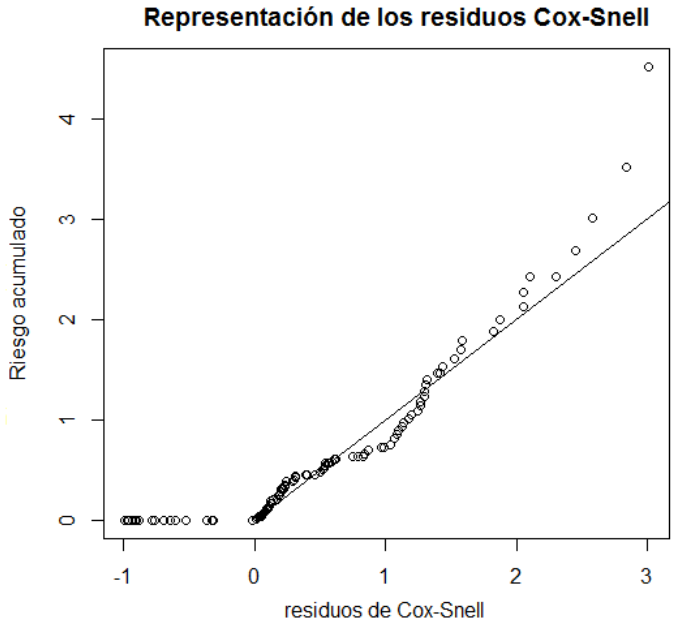


Figura 4.23: Residuos de Cox-Snell para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI* y sin pacientes con tiempos de supervivencia superiores a 15 años. Se representan los residuos de Cox-Snell junto con una recta de pendiente 1 para evaluar el ajuste.

Comprobación de la influencia sobre cada observación en el modelo: Residuos *dfbeta*

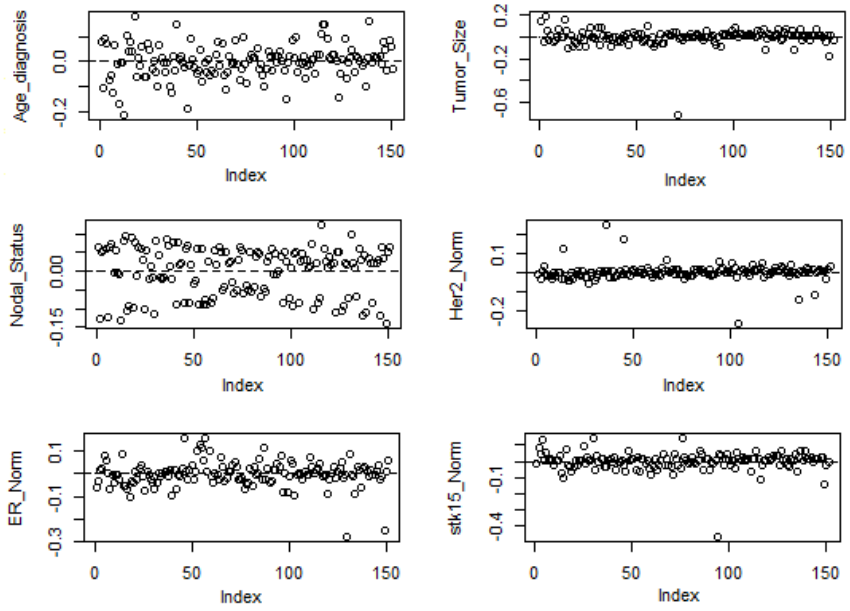


Figura 4.24: Residuos *dfbeta* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*. Para cada covariable, se representa la observación por el cambio de escala aproximada (dividiendo por el error estándar) del coeficiente después de la eliminación de la observación del modelo.

Mediante los residuos *dfbeta* para cada covariable significativa del modelo, Figura 4.24, se observa que no presentan ninguna irregularidad y ningún valor influyente.

Comprobación de la existencia de valores atípicos en el modelo: Residuos de *deviance*

Los residuos de *deviance* obtenidos presentan un patrón definido, esto se debe a que se tienen datos con tiempos de supervivencia superiores a 15 años.

En la Figura 4.25 se representan, sin considerar estos datos, los residuos de *deviance* para este modelo y, se observa que, existe una clara tendencia a sobrevivir más de lo que predice.

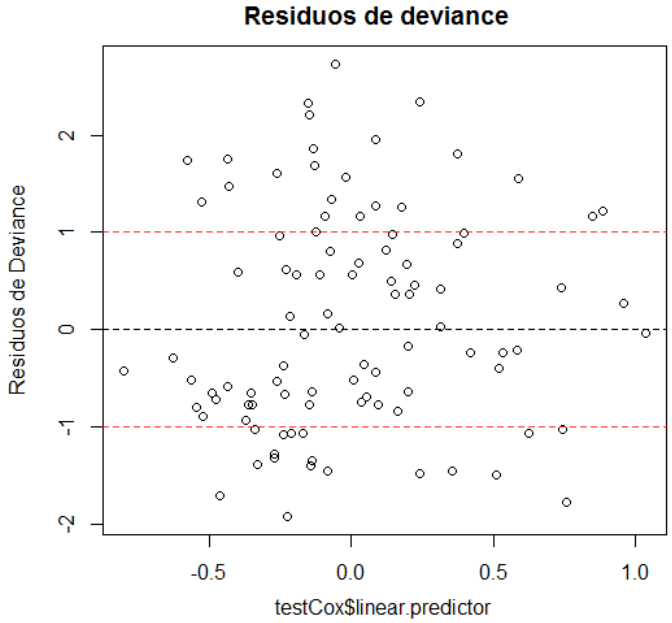


Figura 4.25: Residuos de *deviance* para datos de cáncer de mama en el estudio de supervivencia a 15 años sin *NPI*, para determinar la existencia de *outliers* en el modelo.

Conclusiones

En el estudio sin *NPI* se obtiene que, al eliminar dicha covariable, tanto en la supervivencia a 5 años como a 15 años, cambian las covariables significativas, pues inicialmente entre ellas aparecía *NPI*.

Ahora, en su lugar aparecen *Tumor_Size* y *Nodal_Status*, puesto que *NPI* depende implícitamente del tamaño del tumor, el número de ganglios linfáticos afectados y del grado del tumor.

5

Conclusiones y trabajo futuro

5.1. Conclusiones

Este trabajo afrontaba el reto de desarrollar un análisis de supervivencia para determinar los factores pronóstico para datos reales de cáncer de mama.

Para este fin, se llevó a cabo un estudio de los métodos del marco estadístico teórico, analizando las mejores rutinas que el lenguaje **R** proporciona para ello y desarrollando algunos algoritmos.

Tras este estudio, para comprobar que los algoritmos implementados y las funciones seleccionadas para el análisis son las adecuadas y correctas, se llevó a cabo un *toy-example*.

Una vez verificado que los algoritmos empleados son correctos, se utilizaron datos reales de cáncer de mama proporcionados por la Universidad de Yale, en EEUU [1], para desarrollar el estudio deseado.

En primer lugar, se realizó una estimación de la función de supervivencia mediante el estimador de Kaplan-Meier y, tras la selección de variables significativas, se ajustó el modelo de regresión de Cox para el grupo designado de entrenamiento.

Para verificar el modelo de Cox planteado, se llevaron a cabo los contrastes de hipótesis y el análisis de residuos con el grupo de test.

A la vista de los resultados, se puede concluir que, el cáncer de mama es una enfermedad que aparece fundamentalmente en mujeres con una media de edad de 58,1 años.

La supervivencia a los 5 años es del 70 %, siendo ésta similar en el resto de estudios llevados a cabo sin considerar todas las variables explicativas. La supervivencia a los 10 años es del 62 % frente a un 53 % para los 15 años.

En la selección de variables, para la supervivencia a 5 años, se obtienen como factores de riesgo principales de muerte *Tumor_Size* y *Nodal_Status*, los cuales también aparecen para 10 y 15 años. Además, para la supervivencia a 10 años se obtiene también *Her2_Norm* y para 15 años, *Age_diagnosis*, *Her2_Norm*, *stk15_Norm* y *ER_Norm*.

Tras el análisis de los coeficientes β en el modelo de Cox, se puede concluir que:

- La edad del diagnóstico influye negativamente, ya que aumenta el riesgo de muerte en el estudio de supervivencia a 15 años.
- El tamaño del tumor y el número de ganglios linfáticos afectados son los principales factores de riesgo, ya que aparecen en todos los estudios realizados. El número de ganglios afectados aumenta el riesgo de muerte, de manera considerable, en un factor superior a 1,6 en la mayoría de los casos.
- La presencia de un número elevado de copias del gen *HER2* aumenta el riesgo de muerte para la supervivencia a 10 y 15 años [17].
- El gen *STK15* aumenta el riesgo de muerte solamente para la supervivencia a 15 años [22].
- El receptor de estrógeno, por el contrario, disminuye el riesgo de muerte. Es decir, cuanto mayor sea el número de estrógenos en un paciente, menor riesgo de muerte tendrá [14] [19].

También, con respecto a los resultados obtenidos, se concluye que, el receptor de estrógeno está altamente relacionado con la edad del diagnóstico, ya que, en las mujeres, el número de estrógenos disminuye considerablemente con la edad [23].

En general, el modelo de Cox planteado predice tiempos de supervivencia inferiores a los reales, es decir, existe una tendencia a sobrevivir más de lo que el modelo predice.

Finalmente, el presente trabajo se puede resumir como una introducción al análisis de supervivencia sobre datos de cáncer de mama. En su elaboración, no sólo se ha aprendido a aplicar conceptos matemáticos para la resolución de problemas propios de la ingeniería, sino que también se han adquirido nuevos conocimientos para la extracción automática de información y el manejo de grandes volúmenes de datos. Se han empleado nuevos programas y bibliotecas que hasta ahora no se habían utilizado, se han adquirido conocimientos sobre el campo de la salud y muchas otras competencias que, sin duda, serán útiles a lo largo de la carrera profesional.

5.2. Trabajo futuro

Algunas de las opciones de trabajo futuro planteadas, pero no desarrolladas por la duración finita de este trabajo, son las siguientes:

1. Realizar un estudio más extenso, incluyendo otros métodos y técnicas que no han sido desarrolladas en este trabajo, como puede ser la curva de ROC.

2. El modelo de regresión de Cox desarrollado en este estudio asume que la función de riesgo es constante sobre un periodo de tiempo y el efecto de las covariables se relaciona linealmente con el logaritmo de la razón de riesgos. Si los supuestos del modelo no se cumplen, el modelo de Cox no es el más adecuado.

   Como trabajo futuro sería recomendable buscar alternativas en las que se tuviera en cuenta la dependencia temporal de las covariables. Se sugiere el estudio de aproximaciones mediante *splines*, modelos log-logísticos o el uso de redes neuronales. Cabe destacar que la dificultad en la aplicación de modelos no lineales reside en el tratamiento de datos censurados.

3. La ampliación del estudio a otros campos, como a la teoría de la fiabilidad o experimentos industriales, en los cuales, en lugar del modelo de regresión de Cox, se utiliza el modelo de tiempo de vida acelerada o comúnmente conocido como AFT, el cual supone que las covariables actúan directamente sobre el tiempo de supervivencia.

Glosario de acrónimos

- **AuroraB**: Aurora kinase B
- **AFT**: Accelerated Failure Time model
- **ER**: Estrogen Receptors
- **GRB7**: Growth factor Receptor-Bound protein 7
- **HER2**: Human Epidermal growth factor Receptor 2
- **NPI**: Nottingham Prognostic Index
- **PR**: Progesterone Receptors
- **STK15**: Serine/Threonine Kinase 15

Bibliografía

[1] Yale University. URL: http://www.yale.edu/.

[2] John Fox. Sociology 761. *Introduction to Survival Analysis*, 2006.

[3] Paul D. Allison. *Survival Analysis Using the SAS System.* SAS Institute, 2 edition, 2010.

[4] Frank E. Harrel. *Regression Modeling Strategies.* Springer Series in Statistics, 2001.

[5] David W. Hosmer and Stanley Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Wiley Series in Probability and Statistics, 1998.

[6] D. R. Cox. *Partial likelihood.* Biometrika, 1975.

[7] N. E. Breslow. *Covariance analysis of censored survival data.* Biometrics, 1974.

[8] B. Efron. *The efficiency of Cox's likelihood function for censored data.* Journal of the American Statistical Association, 1977.

[9] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, 1997.

[10] T.M. Therneau, P.M. Grambsch and T.R. Fleming. *Martingale-based residuals for survival models.* Biometrika, 1990.

[11] D. R. Cox and E. J. Snell. *A general definition of residuals.* J.R. Statist. Soc., 1968.

[12] David Collet. *Modelling Survival Data in Medical Research.* Chapman-Hall, 2 edition, 1994.

[13] Cancer Statistics Workig Group. URL: http://www.cdc.gov/uscs/.

[14] American Cancer Society. URL: http://www.cancer.org.

[15] Ministerio de Sanidad-Consumo Área de Epidemiología Ambiental-Cáncer, Instituto Carlos III. *La situación del cáncer en España.* 2005.

[16] MH. Galea, RW. Blamey and IO. Ellis. *The Nottingham Prognostic Index in primary breast cancer.* Breast Cancer Research and Treatment, 1992.

[17] Z. Mitri, T. Constantine and R. O'Regan. *The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy.* 2012.

[18] HS. Feigelson, C. Rodriguez and EJ. Jacobs. *No association between the progesterone receptor gene +331G/A polymorphism and breast cancer.* 2004.

[19] BJ. Deroo and KS. Korach. *Estrogen receptors and human disease.* 2006.

[20] R. Sorrentino, S. Libertini et al. *Aurora B overexpression associates with the thyroid carcinoma undifferentiated phenotype and is required for thyroid carcinoma cell proliferation.* 2005.

[21] Kathleen M. Egan, Polly A. Newcomb et al. *STK15 polymorphism and breast cancer risk in a population-based study.* 2004.

[22] Oregon Health and Science University. *Importance of gene in breast cancer prognosis isolated: GRB7 gene drives an aggressive form of the disease.* 2010.

[23] Medline Plus. URL: http://www.nlm.nih.gov/medlineplus/.

A

Detalle de la biblioteca utilizada

La biblioteca *survival* permite llevar a cabo un análisis de supervivencia para datos que presentan censura. En este anexo, se detallan algunas de las rutinas empleadas para el estudio de supervivencia.

Las rutinas ya existentes en la biblioteca son:

- **Función *Surv***

  La función *Surv* permite crear objetos de supervivencia, los cuales son una estructura de datos que combinan información de tiempo y censura. La estructura de esta función es la siguiente

  ```
  Surv(time, event)
  ```

  El argumento *time* corresponde al tiempo en que el sujeto entra en el estudio y *event* es una variable binaria que indica el estado de censura, considerada 0 si el dato es censurado y 1 si ocurre el evento.

- **Función *survfit***

  La función *survfit* permite crear curvas de supervivencia utilizando el método de Kaplan-Meier. También permite predecir la función de supervivencia para modelos de Cox. La estructura de esta función es

  ```
  survfit(formula)
  ```

  Siendo el argumento *formula* un objeto de supervivencia.

  La función retorna, en forma de tabla, información sobre el número de individuos y eventos en el estudio o el tiempo medio antes de que se presente el evento con respecto a la curva de la función de supervivencia estimada (esto es, el tiempo t tal que $S(t) = 0{,}5$).

- **Función *survdiff***

  Esta función permite efectuar contrastes de hipótesis para verificar la igualdad o diferencia de dos o más curvas de supervivencia. La estructura de la función *survdiff* es la siguiente

  ```
  survdiff(formula)
  ```

  Como antes, el argumento *formula* es un objeto de supervivencia.

- **Función *coxph***

  La función *coxph* permite ajustar los modelos de regresión de Cox.

  Esta función en su forma más sencilla, requiere un objeto de supervivencia y la información de las covariables de cada individuo. Dicha información es ordenada de forma específica y es denotada por *formula* en el argumento de la misma

  ```
  coxph(formula)
  ```

  El argumento opcional de esta función, utilizado para este trabajo, corresponde al método de estimación de las funciones de supervivencia de línea base, explicado en la sección 2.6.1.

  Por un lado, el método de *Breslow* estima el riesgo acumulado base mediante una función no decreciente, calculando así la supervivencia base mediante la relación $supervivencia = \exp[-riesgo\ \ acumulado]$.

  Por otro lado, el método de *Efron* requiere un proceso de cálculo más extenso y se reduce al método de Breslow cuando no hay ningún empate.

  Esta función retorna, mediante una tabla, información acerca de las pruebas locales para verificar que cada coeficiente es significativamente distinto de cero.

  Las columnas de la tabla son información para cada covariable, de manera que, se tienen valores como el coeficiente de regresión estimado, la función exponencial evaluada en el coeficiente, el error estándar del coeficiente de regresión estimado o el p-value que corresponde a dos veces el área acumulada a la derecha del cuantil en una distribución normal con media cero y varianza uno.

  Además, proporciona información de los coeficientes de regresión estimados así como los intervalos de confianza del 95 %.

  También, aparece información resultante tras probar la hipótesis nula de que el vector de variables del modelo son cero, es decir, $H_0 : \beta = \bar{0}$.

- **Función *cox.zph***

  La función *cox.zph* permite llevar a cabo el contraste de hipótesis del modelo de regresión de Cox, considerando como hipótesis nula el cumplimiento del supuesto del modelo de Cox. La estructura de la función es

```
cox.zph(fit)
```

Siendo *fit* el resultado de ajustar el modelo de Cox mediante la rutina *coxph*.

- **Función *residuals***

  Esta función, o en su formato más corto *resid*, es una rutina asociada a los objetos de tipo *coxph*. Permite calcular los residuos de martingala, de puntajes (*score*), de tipo desvío (*deviance*) y de Schoenfeld. Tiene la siguiente estructura

```
residuals(object, type)
```

  Donde *object* es un objeto *coxph* y *type* puede ser «*martingale*», «*deviance*», «*score*» o «*schoenfeld*».

Para la representación de los residuos a partir de los datos se han implementado diferentes algoritmos, los cuales se incluyen a continuación.

```
############################################################
####################ANALISIS DE REDIDUOS####################
############################################################

### Residuos Cox-Snell ###

estado <- test$Censor_5y
mres <- residuals(testCox, type="martingale")
csresi <- estado-mres
hazard.csresi <- survfit(Surv(csresi, estado) ~ 1, type="fleming-harrington")
plot(hazard.csresi$time, -log(hazard.csresi$surv), xlab="residuos de
Cox-Snell", ylab="Riesgo acumulado", lty=1:4, main="Representación de
los residuos Cox-Snell")
lines(c(0,5), c(0,5))

### Residuos dfbeta ###

dfbeta <- residuals(testCox, type="dfbetas")
par(mfrow=c(2,2))
for (j in 1:4){
plot(dfbeta[,j], ylab=names(coef(testCox))[j])
abline(h=0, lty=2, col='black')
lines(c(0,0), c(0,0))
```

```
}

### Residuos de deviance ###

devresi <- resid(testCox, type="deviance")
plot(testCox$linear.predictor, devresi, ylab="Residuos de Deviance",
main="Residuos de deviance")
abline(h=0, lty=2, col='black')
```